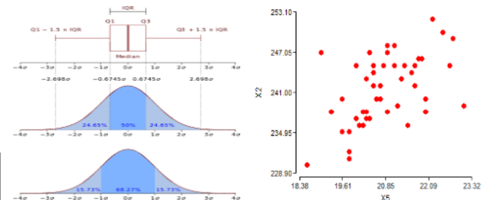
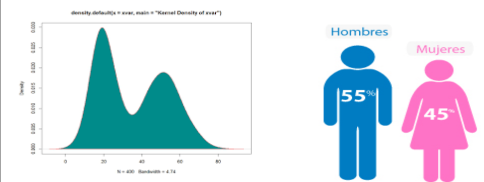






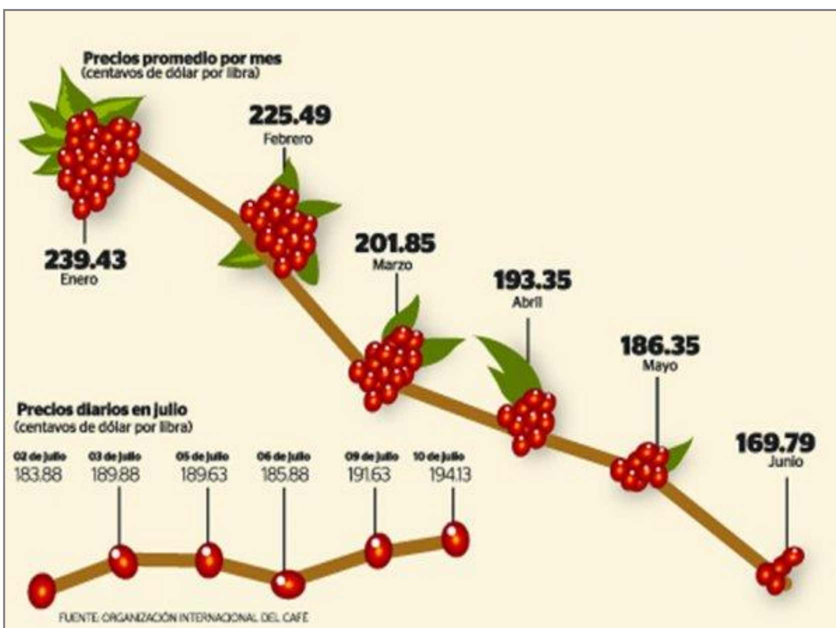


Facultad de Agronomía –Universidad de San Carlos de Guatemala

ESTADÍSTICA

Fundamentos y Aplicaciones en Agronomía y ciencias afines







Ezequiel López

Byron González

2017



Subárea de Métodos de Cuantificación e Investigación
Facultad de Agronomía
Universidad de San Carlos de Guatemala
Oficina C-20, Edificio T-8
Ciudad Universitaria zona 12
Ciudad de Guatemala, C.A.

Versión digital y bases de datos disponibles en: http://cete.fausac.gt

Ciudad de Guatemala, agosto de 2017.





Autores



Ezequiel López es actualmente Profesor Titular VI y Director de la Escuela de Estudios de Postgrado de la Facultad de Agronomía de la Universidad de San Carlos de Guatemala (FAUSAC). Agrónomo egresado de la Escuela Nacional Central de Agricultura –ENCA- (1991), Ingeniero Agrónomo en Sistemas de Producción Agrícola graduado en la FAUSAC (1999), realizó estudios de Maestría en Agronomía (2002) y Doctorado en Ciencias (2014), ambos con Área de concentración en Estadística y Experimentación Agronómica en la Escuela Superior de Agricultura "Luiz de Queiroz" (Piracicaba, SP) de la Universidad de São Paulo, Brasil (USP). Se ha desempeñado como docente de Estadística Aplicada en Agronomía y ciencias afines, en pre y postgrado en la USAC, y en diferentes Universidades Privadas de Guatemala. Además ha sido docente asesor de EPSA, asesor de trabajos de investigación y consultor estadístico en proyectos de investigación. Su área de interés incluye: métodos estadísticos aplicados en Agronomía y ciencias afines, diseño y análisis de experimentos agronómicos, métodos estadísticos multivariados, análisis de regresión, geoestadística y aplicación de modelos lineales mixtos en experimentación agronómica.



Byron González es actualmente Profesor Titular VII, Coordinador de la Sub área de Métodos de Cuantificación e Investigación y Director del Centro de Telemática (CETE) de la Facultad de Agronomía, USAC. Es Doctor en Investigación (2016) por la Universidad Panamericana de Guatemala, realizó estudios de Maestría en Administración de Empresas en la Universidad Rafael Landívar de Guatemala, Ingeniero Agrónomo en Sistemas de Producción Agrícola graduado en la Facultad de Agronomía de la Universidad de San Carlos de Guatemala (1999), Agrónomo egresado de la Escuela Nacional Central de Agricultura (ENCA) 1991. Se ha desempeñado como docente de Informática, Estadística Aplicada en Administración, Economía y Agronomía, en pregrado en la FAUSAC y postgrado en diversas Universidades Privadas de Guatemala. Además como asesor de trabajos de investigación y consultor estadístico en proyectos de investigación. Su área de interés incluye: diseño y análisis de experimentos aplicados en Agronomía y ciencias afines, métodos estadísticos aplicados en el Control de la Calidad, y modelación de regresión.

Consultas y sugerencias:

e-mails: lopez_ezequiel@usac.edu.gt, ealbautis@gmail.com
byrong@usac.edu.gt, byron.gonzalez@cetegt.org

PRESENTACIÓN

El ciudadano común piensa que la Estadística se resume apenas en presentar tablas de números en columnas deportivas o económicas de los periódicos y revistas, ilustradas con gráficos, infografías, esquemas, etc., o a lo más asocian la Estadística a la previsión de resultados electorales. Pero el estadístico de hoy no se limita a compilar tablas de datos e ilustrarlos gráficamente, pues a partir de 1925, con los trabajos de Sir Ronald Fisher, la Estadística se inició como método científico, entonces, el trabajo del estadístico pasó a ser el de ayudar a planear experimentos, interpretar y analizar los datos experimentales y presentar los resultados a manera de facilitar la toma de decisiones razonables. De este modo, se puede definir Estadística como la ciencia derivada de las matemáticas que se preocupa de la recolección, organización, presentación, análisis e interpretación de datos, así como de hacer inferencias.

En estas notas de acompañamiento, se describen de manera teórica y práctica los temas contenidos en el programa del curso de Estadística General que se brinda en las carreras de Ingeniería Agronómica, Ingeniería Forestal e Ingeniería Ambiental de Guatemala. No se pretende, por supuesto, hacer alguna aportación novedosa a la copiosa literatura ya existente, sino que la idea fundamental es la de recopilar e integrar en un documento los contenidos temáticos (unidades, temas y subtemas), que marca el programa de dicha materia, mismos que se encuentran dispersos en la bibliografía manejada en los cursos de Estadística. Estas notas abarcan unidades referentes a: conceptos generales, estadística descriptiva, introducción al estudio de las probabilidades, distribuciones de probabilidad (discretas y continuas), inferencia estadística (estimación y pruebas de hipótesis), análisis de correlación lineal simple, análisis de regresión lineal simple y múltiple.

Este documento viene a llenar un vacío enorme en la enseñanza de la Estadística en el campo agronómico de Guatemala, ya que en nuestro medio es difícil conseguir literatura específica para esta área. Por medio de ejemplos prácticos, extraídos de trabajos de investigación realizados en el campo agronómico y forestal de Guatemala y algunos países latinoamericanos (Argentina, Brasil, Chile, Colombia, México y Perú), se ilustra la importancia que tiene la Estadística, como herramienta de apoyo en otras materias, tales como: Dasometría, Economía, Edafología, Entomología, Fitopatología, Hidrología, Inventarios Forestales, Genética, entre otras. Además se incluyen aplicaciones en Infostat v. 2015 y aportes de nuestra experiencia en el postgrado y durante los años que nos hemos desempeñado como catedráticos de los cursos de Estadística en la Facultad de Agronomía de la Universidad de San Carlos de Guatemala y en diferentes Universidades privadas del país.

Deseamos patentar nuestro agradecimiento a los profesores de Estadística, de la Facultad de Agronomía (USAC): Ing. Agr. M.C. Víctor Álvarez Cajas (QEPD) e Ing. Agr. M.Sc. Marino Barrientos; del postgrado en Estadística de la ESALQ (USP): Dra. Clarice García Borges Demetrio y Dr. Sc. Carlos Tadeu dos Santos Dias. A los colegas docentes de la carrera de Ingeniería Forestal del campus “San Pedro Claver” (URL/Cobán): Ing. MBA Carlos Archila, Ing. M.Sc. Estuardo Vaides, Ing. M.Sc. Ricardo Ávila, Ing. Roberto Moya e Ing. M.Sc. Manuel Sabino Mollinedo, así como al Ing. MA. Carlos Ardón y P. Agr. Francisco Pec, docentes de la FAUSAC, por las observaciones al utilizar este texto. Y a los estudiantes que han utilizado las versiones iniciales de este documento, que con sus dudas dentro y fuera del aula, han contribuido a mejorar este texto. Finalmente, esperamos sus comentarios y sugerencias para ir mejorando cada día estas notas de acompañamiento.

Guatemala agosto de 2017.

CONTENIDO

UNIDAD I
ASPECTOS INTRODUCTORIOS

1.1 DEFINICIÓN DE ESTADÍSTICA

Walker (1929) atribuye el primer uso del término “estadística” al profesor alemán, Gottfried Achenwall (1719 – 1772), quien utilizó la palabra alemana *Statistik*, que extrajo del término italiano *statista* (estadista). Creía, y con sobrada razón, que la nueva ciencia sería el aliado más eficaz del gobernante consciente, para la planificación de los recursos. La raíz de la palabra se halla, por otra parte, en el término latino *status*, que significa estado o situación. Indicando con esto la importancia histórica de la recolección de datos por parte del gobierno de un país, relacionados principalmente a información demográfica (censos por ejemplo).

El Dr. E. A. W. Zimmerman introdujo el término statistics (estadística) a Inglaterra. Su uso fue popularizado por Sir John Sinclair (1754 – 1835) en su obra *Statistical Account of Scotland* 1791 – 1799 (“Informe estadístico sobre Escocia 1791 – 1799”). Sin embargo mucho antes del siglo XVII, la gente ya la utilizaba y registraba datos. A continuación se presentan algunas definiciones de Estadística:

a) Ciencia derivada de la matemática que se ocupa de la extracción de la información contenida en datos provenientes de muestras y de su uso para hacer inferencias acerca de la población de donde fueron extraídos estos datos.

b) La Estadística estudia los métodos científicos para recolectar, organizar, resumir y analizar datos, así como para extraer conclusiones válidas y tomar decisiones razonables basadas en tal análisis.

1.2 HISTORIA DE LA ESTADÍSTICA

Los comienzos de la estadística pueden ser hallados en el antiguo Egipto, donde los faraones lograron recopilar, alrededor del año 3050 antes de Cristo, datos relativos a la población y la riqueza del país. De acuerdo al historiador griego Heródoto, este registro de riqueza y de población se hizo con el objetivo de preparar la construcción de las pirámides. En el mismo Egipto, Ramsés II hizo un censo de las tierras con el objeto de verificar un nuevo reparto.

En el antiguo Israel, la Biblia da referencias en el libro de los Números, de los datos estadísticos obtenidos en dos recuentos de la población hebrea. El rey David por otra parte, ordenó a Joab, general del ejército hacer un censo de Israel con la finalidad de conocer el número de la población. Los chinos efectuaron censos hace más de cuarenta siglos y los griegos efectuaron censos periódicamente con fines tributarios, sociales (división de tierras) y militares (cálculo de recursos y hombres disponibles).

Pero fueron los romanos, maestros de la organización política, quienes mejor supieron emplear los recursos de la Estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos tenían la obligación de anotar nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas. Para el nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del imperio.

Durante los mil años siguientes a la caída del imperio Romano se realizaron muy pocas operaciones estadísticas, con la notable excepción de las relaciones de tierras pertenecientes a la Iglesia, compiladas por Pipino el Breve en el 758 y por Carlomagno en el 762 DC. Durante el siglo IX se realizaron en Francia algunos censos parciales de siervos. En Inglaterra, Guillermo el Conquistador recopiló el *Domesday Book* o libro del Gran Catastro para el año 1086, un documento de la propiedad, extensión y valor de las tierras de Inglaterra. Esa obra fue el primer compendio estadístico de Inglaterra.

Aunque Carlomagno en Francia y Guillermo el Conquistador en Inglaterra, trataron de revivir la técnica romana, los métodos estadísticos permanecieron casi olvidados durante la Edad Media. Durante los siglos XV, XVI, y XVII, Leonardo de Vinci, Nicolás Copérnico, Galileo, Neper, William Harvey, Sir Francis Bacon y René Descartes, hicieron grandes contribuciones al método científico, de tal forma que cuando se crearon los Estados Nacionales y surgió como fuerza el comercio internacional, existía ya un método capaz de aplicarse a los datos económicos.

Para el año 1532 empezaron a registrarse en Inglaterra las defunciones debido al temor que Enrique VII tenía por la peste. Más o menos por la misma época, en Francia la ley exigió a los clérigos registrar los bautismos, fallecimientos y matrimonios. Durante un brote de peste que apareció a fines de la década de 1500, el gobierno inglés comenzó a publicar estadísticas semanales de los decesos. Esa costumbre continuó muchos años, y en 1632 estos *Bills of Mortality* (Cuentas de Mortalidad) contenían los nacimientos y fallecimientos por sexo. En 1662, el capitán John Graunt usó documentos que abarcaban treinta años y efectuó predicciones sobre el número de personas que morirían de varias enfermedades y sobre las proporciones de nacimientos de varones y mujeres que cabría esperar.

El primer empleo de los datos estadísticos para fines ajenos a la política tuvo lugar en 1691 y estuvo a cargo de Gaspar Neumann, un profesor alemán que vivía en Breslau. Este investigador se propuso destruir la antigua creencia popular de que en los años terminados en siete moría más gente que en los restantes, y para lograrlo hurgó pacientemente en los archivos parroquiales de la ciudad. Después de revisar miles de partidas de defunción pudo demostrar que en tales años no fallecían más personas que en los demás. Los procedimientos de Neumann fueron conocidos por el astrónomo inglés Halley, descubridor del cometa que lleva su nombre, quien los aplicó al estudio de la vida humana. Sus cálculos sirvieron de base para las tablas de mortalidad que hoy utilizan todas las compañías de seguros.

Durante el siglo XVII y principios del XVIII, matemáticos como Bernoulli, Francis Maseres, Lagrange y Laplace desarrollaron la teoría de probabilidades. No obstante durante cierto tiempo, la teoría de las probabilidades limitó su aplicación a los juegos de azar y hasta el siglo XVIII no comenzó a aplicarse a los grandes problemas científicos. Thomas Bayes (Londres, Inglaterra, 1702 - Tunbridge Wells, 1761), fue uno de los primeros en utilizar la probabilidad inductivamente y establecer una base matemática para la inferencia probabilística. Actualmente, con base en su obra, se ha desarrollado una poderosa teoría que ha conseguido notables aplicaciones en las más diversas áreas del conocimiento.

Godofredo Achenwall, profesor de la Universidad de Gotinga, acuñó en 1760 la palabra estadística. Jacques Quételect es quien aplica la Estadística a las ciencias sociales. Él interpretó la teoría de la probabilidad para su uso en las ciencias sociales y resolver la aplicación del principio de promedios y de la variabilidad a los fenómenos sociales. Entretanto, en el período del 1800 al 1820 se desarrollaron dos conceptos matemáticos fundamentales para la teoría estadística; la teoría de los errores de observación, aportada por Laplace y Gauss; y la teoría de los mínimos cuadrados desarrollada por Laplace, Gauss y Legendre. A finales del siglo XIX, Sir Francis Galton dio forma al método conocido como regresión. De aquí partió el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica como J. Pease Norton, R. H. Hooker y G. Udny Yule, que efectuaron amplios estudios sobre la medida de las relaciones. Más adelante, a partir de 1919 la estadística experimental tuvo su desarrollo cuando Ronald A. Fisher asumió la dirección del departamento de Estadística de la Estación

Experimental de Rothampstead en Londres, Inglaterra, así mismo son importantes los trabajos sobre análisis de la varianza y diseño y análisis de experimentos, desarrollados por el matemático estadounidense George Snedecor.

Hoy el uso de la estadística se ha extendido más allá de sus orígenes como un servicio al estado o al gobierno. Personas y organizaciones usan estadística para entender datos y tomar decisiones en ciencias naturales y sociales, medicina, negocios y otras áreas. La estadística es pensada generalmente no como una subárea de las matemáticas sino como una ciencia diferente "aliada". Muchas universidades tienen departamentos en matemáticas y estadística separadamente. La estadística es enseñada en departamentos tan diversos como psicología, educación, agronomía, ciencias forestales, ingeniería, economía y salud pública. Y con el desarrollo tecnológico de la computación, la estadística ha avanzado a pasos agigantados, mejorando cada día sus procedimientos de cálculo.

En Guatemala, de acuerdo con la información del Instituto Nacional de Estadística disponible en su sitio web (https://www.ine.gob.gt/index.php/about/historia), el primer testimonio de la actividad estadística data de 1778, cuando se realizó el Primer Censo de Población levantado por las autoridades eclesiásticas que incluían los registros parroquiales de nacimientos, matrimonios y defunciones.

A partir de 1821 se dieron las primeras inquietudes por organizar la estadística oficial. José Cecilio del Valle figura como primer intelectual preocupado por la estructuración estadística en el istmo centroamericano, por lo cual es considerado “El Padre de la Estadística” en Guatemala. Escribió artículos como “La Estadística Plataforma del Enaltecimiento Social”, publicados en el periódico “El Amigo de la Patria”, en los que insistió sobre la importancia de la Estadística y su campo de aplicación. Su esfuerzo se materializó con la promulgación de la ley sobre la formación de la estadística en las provincias de Centro América el 15 de noviembre de 1823, reconociéndola como la Primera Ley Estadística.

El 19 de mayo de 1824, por decreto se ordena el levantamiento de los censos de población. El 13 de julio de 1825 fue conformada la Primera Comisión Nacional de Estadística, la cual marcó el inicio de la recolección de información para la gestión de gobierno incluyendo la elaboración de las primeras nóminas y establecimientos comerciales e industriales. La Sección de Estadística fue fundada en 1879, adscrita al Ministerio de Fomento, con carácter de Oficina Central de Estadística que se encargó en 1880 de levantar el Segundo Censo de Población del país. En 1886, esta Oficina fue elevada a la categoría de Dirección General de Estadística, llevando a cabo en 1893 y en 1921 el III y IV Censos Generales de Población respectivamente.

En 1936, la Dirección General de Estadística se incorpora al Ministerio de Hacienda y en marzo del mismo año se promulga la Segunda Ley Estadística. En 1938 se levantó el Censo Urbano de la Ciudad de Guatemala. El V Censo de Población se realizó en 1940. La Dirección de Estadística pasa a jurisdicción del Ministerio de Economía y Trabajo en 1944. En 1950 se llevan a cabo el VI Censo de Población, I Censo Agropecuario y el I Censo de Vivienda Urbana. En 1958 se emite la Tercera Ley Estadística, que creó el Sistema Estadístico Nacional. Esta Ley estuvo vigente hasta 1985, cuando fue promulgada la Ley Orgánica del Instituto Nacional de Estadística, (Decreto Ley 3-85) convirtiéndose la institución en un ente descentralizado y semiautónomo.

En el ámbito académico, la subárea de Métodos de Cuantificación e Investigación y el Centro de Telemática de la Facultad de Agronomía han estado a la vanguardia en el desarrollo y divulgación de la Estadística y software estadístico, con el apoyo de la Sociedad Internacional de Biometría (IBS), Región Centro América y el Caribe y de la USAC, se han organizado eventos nacionales, con la participación de reconocidos estadísticos latinoamericanos y generado materiales de apoyo para la docencia. Así mismo se han ido iniciando trabajos de investigación en áreas como: geoestadística y análisis multivariado de datos.

1.3 DIVISIÓN DE LA ESTADÍSTICA

La Estadística para su mejor estudio se ha dividido en tres grandes ramas: Estadística Descriptiva, Probabilidades y la Estadística Inferencial.

1.3.1 La Estadística Descriptiva consiste en la presentación de datos en forma numérica, tablas y gráficas. Esta comprende cualquier actividad relacionada con los datos y está diseñada para resumir o describir los mismos, sin factores pertinentes adicionales; esto es, sin intentar inferir nada que vaya más allá de los datos, como tales. Es en general utilizada en la etapa inicial de los análisis, cuando se tiene contacto con los datos por primera vez.

1.3.2 La Probabilidad puede ser pensada como la teoría matemática utilizada para estudiar la incertidumbre oriunda de fenómenos de carácter aleatorio, o sea, producto del azar.

1.3.3 La Estadística Inferencial se deriva de muestras, de observaciones hechas sólo acerca de una parte de un conjunto numeroso de elementos y esto implica que su análisis requiere de generalizaciones que van más allá de los datos. Como consecuencia, la característica más importante del reciente crecimiento de la estadística ha sido un cambio en el énfasis de los métodos que describen a métodos que sirven para hacer generalizaciones. La Estadística Inferencial investiga o analiza una población partiendo de una muestra tomada.

La Estadística descriptiva y la inferencial comprenden la estadística aplicada. Hay también una disciplina llamada estadística matemática, la cual se refiere a las bases teóricas de la materia, e incluye el estudio de las probabilidades.

Otra división de la estadística es:

1.3.4 Estadística Paramétrica: en este tipo de estadística el interés es hacer estimaciones y pruebas acerca de uno o más parámetros de la población. Además, en todas estas estimaciones y pruebas de hipótesis se establece como suposición general que la población o poblaciones de donde provienen las muestras deben estar distribuidas normalmente, aunque sea en forma aproximada.

13.5 Estadística No Paramétrica (o de distribución libre): estudia las pruebas y modelos estadísticos cuya distribución subyacente no se ajusta a los llamados criterios paramétricos. Su distribución no puede ser definida *a priori*, pues son los datos observados los que la determinan. La utilización de estos métodos se hace recomendable cuando no se puede asumir que los datos se ajusten a una distribución normal o cuando el nivel de medida empleado no sea, como mínimo, de intervalo.

Otras ramas importantes de la Estadística son:

Geoestadística: comprende a un conjunto de herramientas y técnicas que sirven para analizar y predecir los valores de una variable que se muestra distribuida en el espacio o en el tiempo de una forma continua. Debido a su aplicación orientada a los Sistemas de Información Geográfica (SIG), también se podría definir como la estadística relacionada con los datos geográficos.

Inferencia Bayesiana: la metodología bayesiana está basada en la interpretación subjetiva de la probabilidad y tiene como punto central el Teorema de Bayes. Los modelos bayesianos primordialmente incorporan conocimiento previo para poder estimar modelos útiles dentro de un espacio muestral y de este modo poder estimar parámetros que provengan de la experiencia o de una teoría probabilística.

Estadística Multivariada: las técnicas estadísticas multivariadas permiten establecer, a partir de numerosos datos y variables ciertas relaciones, investigar estructuras latentes y ensayar diversas maneras de organizar dichos datos, bien transformándolos y presentándolos bajo una forma nueva más asequible, bien reduciéndolos, sin perder demasiada información inicial, hasta componer un resumen lo más completo posible del conjunto de datos original, habitualmente bastante complejo.

1.4 INDIVIDUO O UNIDAD ESTADÍSTICA

Todo estudio estadístico se hace sobre un **individuo**, que es el objeto de observación. Una unidad estadística es la entidad sobre la que se quiere obtener los datos para ser analizados. Por ejemplo, una unidad estadística puede ser una persona adulta, un saco con 45 kg de café, un escritorio, un árbol de cedro, una parcela de 50 m^2 con tomate, un río, una vaca, un tallo o una macolla de caña de azúcar, una colonia de hongos, un tractor, etc.

1.5 POBLACIÓN

La población es el conjunto de todas las unidades estadísticas. También se puede definir como el conjunto de individuos que tienen por lo menos una característica en común observable. El estudio completo de una población se denomina: **Censo**. En noviembre de 2002 se realizó en Guatemala el XI Censo Nacional de Población y el VI Censo Nacional de Habitación cuyos resultados fueron presentados en febrero del 2003. En mayo del 2003 se realizó el IV Censo Nacional Agropecuario. Para su estudio, en general se clasifican en poblaciones finitas (constan de un número determinado de elementos, susceptible a ser contado) y poblaciones infinitas (tienen un número indeterminado de elementos, los cuales no pueden ser contados).

1.6 MUESTRA

Es el subconjunto de una población, que manifiesta las mismas características de la población original de donde fue extraída. Los requisitos deseables de una buena muestra son: representatividad y confiabilidad. Lo primero se consigue a través de la selección del tipo de muestreo adecuado, en tanto que la confiabilidad está referida al tamaño de la muestra.

Los estudios que involucran la toma de muestras se denominan: **Encuestas.** En 1986/87 se inicia en Guatemala el Sistema Nacional de Encuestas de Hogares. En 1995/98 se realiza la Encuesta Nacional Salud Materno Infantil (ENSMI). En 1998–1999 se realiza la Segunda Encuesta Nacional de Ingresos y Gastos Familiares – ENIGFAM – y es ésta la que sienta las bases para el nuevo índice de precios al consumidor del año 2000, IPC actual.

En el año 2000 la Encuesta Nacional sobre Condiciones de Vida de los Hogares, ENCOVI 2000 amplía el universo de información que el INE ofrece a los usuarios. Con el propósito de proveer al Sistema Estadístico Nacional de información actualizada del sector agropecuario, el Instituto Nacional de Estadística de Guatemala realizó la Encuesta Nacional Agropecuaria (ENA) en el año 2006. Esta encuesta tuvo como objetivo la obtención de estimaciones de la producción agropecuaria y existencia de ganado a nivel nacional y departamental. Dentro de las actividades recientes que ha realizado el INE están: la Encuesta Nutricional 2012, la Encuesta Nacional de Empleo e Ingresos -ENEI 1-2014-, la Encuesta Nacional Agropecuaria 2013 y las Estadísticas de empleo y trabajo no remunerado de las mujeres en Guatemala, entre otras. Para mayor información sobre estas encuestas y censos realizados en Guatemala, consulte el sitio oficial del Instituto Nacional de Estadística (INE): www.ine.gob.gt.

1.7 PARÁMETRO

Es un valor que resume la información de una población, se denota con letras del alfabeto griego. Por ejemplo: μ = media, σ = desviación estándar, σ^2 = varianza, ρ = coeficiente de correlación poblacional. Los parámetros son obtenidos a partir de mediciones realizadas en los censos.

1.8 ESTIMADOR O ESTADÍSTICO

Es un valor representativo de una muestra, se denota con letras del alfabeto arábigo. Por ejemplo: $\bar{x}$ = media, s = desviación estándar, s^2 = varianza, r = coeficiente de correlación de la muestra.

1.9 VARIABLES Y SU CLASIFICACIÓN

1.9.1 DEFINICIÓN DE VARIABLE

Una variable en estadística es lo que se observa o mide sobre las unidades estadísticas. Son características que varían de un individuo a otro. Las variables son representadas con letras mayúsculas, por ejemplo: X, Y, Z, etc. Y los valores que asumen, son representados con letras minúsculas, ejemplo: *x*, *y*, *z*.

1.9.2 CLASIFICACIÓN DE LAS VARIABLES

Dependiendo de su naturaleza, se distinguen dos tipos principales de variables:

a) CUANTITATIVAS

Son aquellas que expresan cantidades y los resultados son de tipo numérico. Se clasifican en:

a.1 CUANTITATIVAS DISCRETAS

También llamadas de conteo, son las que no aceptan valores decimales. Ejemplo: número de hijos por familia, número de camas de un hospital, cantidad de plantas de café por metro cuadrado, número de áfidos por planta, número de brotes por planta, número de racimos de banano por hectárea, número de ausencias de un trabajador por mes. Matemáticamente se pueden representar de la siguiente manera:

Sea X el número de árboles con cáncer en una muestra de 10 árboles: $X \in \{0, 1, 2, 3, \ldots, 9, 10\}$
Sea Y el número de plántulas en un área de 10 m^2 de suelo en una floresta nativa: $Y \in \{0, 1, 2, , \ldots\}$

a.2 CUANTITATIVAS CONTINUAS

Este tipo de variables pueden tomar como valores cualquier número real, es decir, un valor con decimales, y que resultan de un proceso de medición. Ejemplo: altura de una planta, peso de semillas, temperatura de un cuarto frío, diámetro de un árbol, caudal de un río, precipitación pluvial, etc.

- Sea D el diámetro de árboles de *Pinus maximinoii* en una plantación: $D \in [10, 50]$
- Sea W la biomasa(tm.ha^{-1}) en un ecosistema forestal: $W \in [500, 5000]$

b) Variables CUALITATIVAS

Son las variables que presentan como posibles resultados una cualidad o atributo del individuo investigado. Las posibles cualidades que tiene una variable cualitativa se llaman: modalidades de la variable. Algunos autores también las llaman: Categorías o Atributos. Según el número de categorías, algunos autores las clasifican como: **dicotómicas**, solo hay dos modalidades, en general los fenómenos de respuesta binaria, como: padecer una enfermedad (sí, no), sexo (femenino, masculino), resultado de un prueba de evaluación (aprobado, reprobado); o bien **politómicas**, cuando hay más de dos categorías, por ejemplo: estado civil, color de los ojos de una persona, lugar de origen, profesión, forma de la hoja de una planta, susceptibilidad de una planta a una enfermedad, nivel de satisfacción de un cliente, etc.

1.10 ESCALAS DE MEDICIÓN

1.10.1 INFORMACIÓN CUALITATIVA

a) Escala Nominal

Es la escala más débil en cuanto a la información que proporciona. Como su nombre lo indica, esta escala consiste en "nombrar a las observaciones". Para distinguir los agrupamientos de unidades se emplean símbolos, letras o números. En el caso de que se empleen números, estos solo tienen un carácter simbólico y no numérico. Ejemplo:

- Especies arbóreas presentes en la parte alta de la cuenca del río Naranjo, San Marcos.
- Estado civil de los habitantes del caserío "San Martín", San Martín Sacatepéquez, Quetzaltenango (soltero, casado, divorciado, unido).
- Tipos de uso del suelo (agrícola, forestal, pecuario, etc.) en el municipio de Chiantla.
- Municipio de procedencia de los estudiantes de la carrera de Ingeniería Forestal de la URL.

b) Escala Ordinal

En este nivel, las unidades de los grupos guardan cierta relación entre sí, que se pone de manifiesto cuando se está en posibilidad de establecer una relación de tipo mayor o menor que. Ejemplos:

- Nivel de estudios, ya que sus modalidades están ordenadas según la duración de los estudios: Educación primaria, secundaria, diversificado, universitaria.
- Grado de aceptación de algún producto: buena, regular, mala.
- Nivel socioeconómico de una familia (alto, medio, bajo)

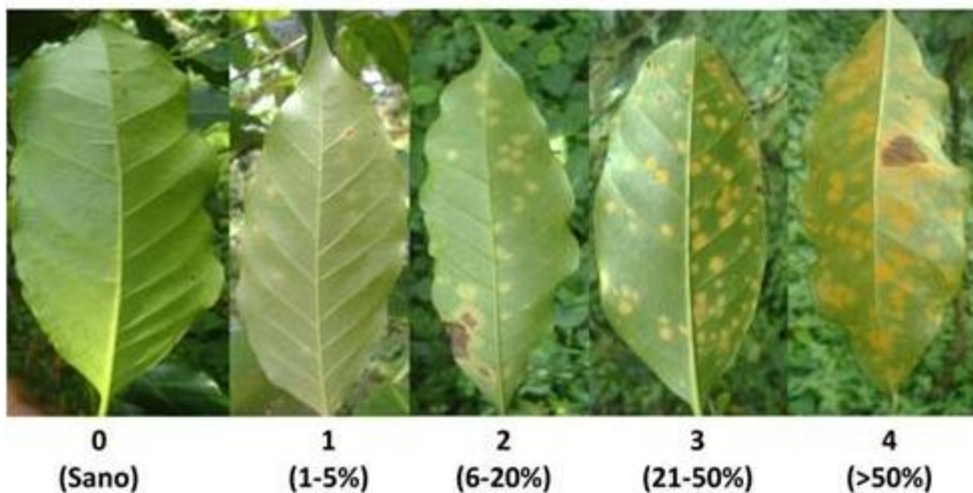
La escala de Likert es un ejemplo claro de escala de medición ordinal. Un elemento de tipo Likert es una declaración que se le hace a los sujetos para que éstos lo evalúen en función de su criterio subjetivo; generalmente se pide a los sujetos que manifiesten su grado de acuerdo o desacuerdo. Normalmente hay 5 posibles respuestas o niveles de acuerdo o desacuerdo, aunque algunos evaluadores prefieren utilizar 7 o 9 niveles. El formato de un típico elemento de Likert con 5 niveles de respuesta sería:

Nivel de satisfacción del servicio que presta la Oficina de Atención al Estudiante:

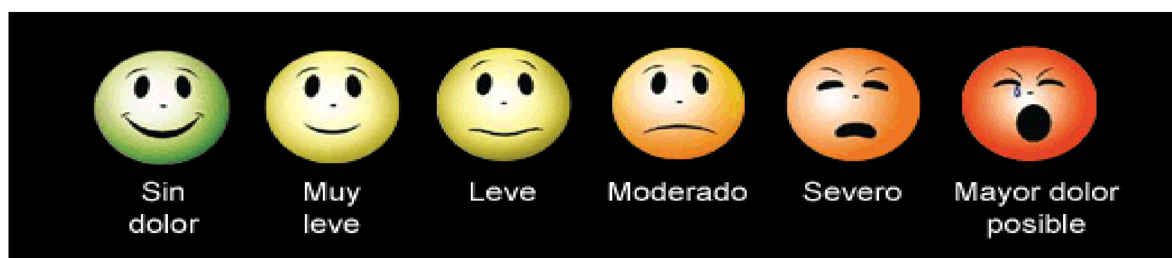
1. Totalmente en desacuerdo
2. En desacuerdo
3. Ni de acuerdo ni en desacuerdo
4. De acuerdo
5. Totalmente de acuerdo.

Otra variable medida en escala ordinal, es la severidad de una enfermedad. Vea la siguiente escala diagramática para realizar la evaluación de severidad de la roya del café en hojas.

CLASES	HOJA (% DAÑO)
0	Sano sin síntomas visibles
1	1-5 % de área afectada
2	6-20 % de área afectada
3	21-50 % de área afectada
4	> 50% de área afectada



Otro ejemplo es la escala de valoración de caras de dolor de Wong-Baker, usada en medicina.



1.10.2 INFORMACIÓN CUANTITATIVA

a) Escala de Intervalo

Este tipo de escala provee información mucho más precisa, a la vez que permite llevar a cabo mediciones mucho más sofisticadas que las escalas nominal u ordinal. La escala de intervalo no sólo informa acerca del orden de unos objetos, sino que también acerca de las distancias o diferencias numéricas entre dichos objetos. De hecho, esta escala permite medir y comparar esas distancias o diferencias con precisión. En otras palabras (y de aquí el nombre de escalas de intervalo), las distancias o 'intervalos' de igual tamaño en la escala son de hecho iguales no importando dónde en la escala se realice la medición. Por ejemplo, los resultados numéricos de los exámenes académicos (rango de 0 a 100) pueden ser medidos usando escalas de intervalo.

La escala de intervalo, sin embargo, no posee una definición única del valor cero. En otras palabras, el cero es arbitrario en el sentido de que no representa ausencia absoluta de la característica que se desea medir. En este sentido las escalas de intervalo son equivalentes a termómetros, en los que el valor cero no representa la ausencia absoluta de calor. En el ejemplo anterior, si un estudiante obtiene un resultado de cero puntos en un examen, ello obviamente no significa que el estudiante no sepa absolutamente nada acerca de la materia evaluada. El comportamiento humano es casi siempre medido utilizando escalas. Otras variables medidas en esta escala son: temperatura, horario meridiano, grados de latitud o de longitud. La numeración de los años en nuestro calendario utiliza también una escala de intervalos. Las autoridades eclesiásticas y gubernamentales de la época decidieron arbitrariamente fijar como el año 1 el del nacimiento de Cristo y como unidad de medida un lapso de 365 días.

b) Escala de Razón

Los atributos son cuantitativos organizados en una escala donde tanto el intervalo entre dos valores, como el punto cero, tienen significado real (indica ausencia de valor). Dadas dos medidas en esta escala, podemos decir si son iguales, o si una es diferente, mayor, que tan mayor y cuantas veces la otra. La altura de un individuo es un ejemplo de la medida en esta escala. Si ella fuera medida en centímetros (cm), 0 cm es el origen y 1 cm es la unidad de medida. Un individuo con 190 cm es dos veces más alto que un individuo con 95 cm, y esta relación continua valiendo si usamos 1 cm como unidad. Otras variables que son medidas en esta escala son: peso, longitud, diámetro, volumen, estatura, densidad.

1.11 INDICADOR

Un indicador es un elemento extraído de la realidad que permite cuantificar ciertas características medibles, y que posteriormente será la base para la conformación de índices relativos de acuerdo con los valores obtenidos. Puede decirse también que son elementos conceptuales que sirven para señalar o indicar que una característica o variable está ocurriendo.

Existen variables cuyos indicadores pueden tener un menor grado de objetividad (como participación política, desintegración familiar, interés por el trabajo comunitario) en comparación con otras variables (por ejemplo: deterioro de la vivienda, concentración de la riqueza, escolaridad, tipo de ocupación, etc.)

Algunos ejemplos se presentan a continuación:

- Indicadores de población:

Población total de un país, por regiones, estados, municipios, etc.
Densidad de habitantes por área.
Población económicamente activa.
Población urbana y rural.
Índice de crecimiento poblacional.

- Indicadores económicos

Producto Interno Bruto (PIB)
Superficie agrícola explotable
Índice nacional de precios al consumidor
Explotación pesquera por especie
Volumen de producción forestal.

- Indicadores del comercio

Número de establecimientos por ramo
Volumen de exportación manufacturada
Comercios registrados en el Ministerio de Finanzas

- Indicadores educativos:

Número de mujeres analfabetas
Porcentaje de deserción en el nivel primario.
Número de egresados de la Universidad.
Porcentaje de reprobación en el nivel básico.
Porcentaje de titulación en el diversificado.
Índice de aprobación de Estadística General.

- Indicadores de desarrollo socioeconómico:

Salario mínimo
Impuestos promedio por habitante
Número de vehículos registrados
Oferta hotelera: número de cuartos.

- Indicadores sociopolíticos

Número de electores
Resultado de votación por partido
Número de sindicatos por sector o actividad.

1.12 NOTACIÓN SUMATORIA

1.12.1 DEFINICIÓN

Como la operación de adición ocurre frecuentemente en Estadística, se utiliza la letra griega Σ (sigma mayúscula) como indicación de: "realizar la suma de . . . . ."

La expresión $\sum_{i=1}^{n} x_i$ significa: "sumatoria de x_i, para i variando de 1 hasta n". En otras palabras:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Ejemplo 1: Sea X la variable aleatoria cuantitativa contínua, peso (expresado en kilogramos) de 10 estudiantes del curso de Estadística General, sección "A" de la Facultad de Agronomía.

I	1	2	3	4	5	6	7	8	9	10
x_i	60.5	55.0	72.8	80.9	55.0	60.0	58.0	47.0	57.8	85.2
	(x_1)	(x_2)	(x_3)	.	.	.	.	.	.	(x_{10})

Resolver:

$$\sum_{i=1}^{10} x_i = x_1 + x_2 + \ldots + x_{10} = 632.2 \text{ kg}$$

$$\sum_{i=3}^{6} x_i = x_3 + x_4 + x_5 + x_6 = 268.7 \text{ kg}$$

En Estadística también estamos frecuentemente interesados en obtener la suma de los valores de una variable al cuadrado, por tanto:

$$\sum_{i=1}^{10} x_i^{\ 2} = 60.5^2 + 55^2 + \ldots + 85.2^2 = 41{,}327.78 \text{ kg}^2$$

Con este resultado se puede concluir que $\sum_{i=1}^{n} x_i^{\ 2}$, o sea, ***la suma de cuadrados***, no es lo mismo que $\left(\sum_{i=1}^{n} x_i\right)^2$, que es conocida como ***el cuadrado de la suma***. Esto es:

$$\sum_{i=1}^{n} x_i^{\ 2} \neq \left(\sum_{i=1}^{n} x_i\right)^2$$

$$41{,}327.78 \neq 399{,}676.84$$

1.12.2 ALGUNAS PROPIEDADES DE LA NOTACIÓN SUMATORIA

1. Una constante sumada *n* veces, será igual a *n* veces el valor de la constante. Ej.: Sea *k* una constante cualquier, entonces:

   $\sum_{i=1}^{n} k = n \times k$, o sea, $\sum_{i=1}^{n} k = k + k + k + \ldots + k = n \times k$

   Por ejemplo: Sí *k* = 3 y *n* = 5, se tiene que: $\sum_{i=1}^{5} 3 = 3 + 3 + 3 + +3 + 3 = 5 \times 3 = 15$

2. La sumatoria del producto de una constante por una variable, es igual a esa constante multiplicada por la sumatoria de la variable.

   $\sum_{i=1}^{n} k \times x_i = k \times \sum_{i=1}^{n} x_i$, donde *k* es una constante.

   Por ejemplo: si se desea convertir los pesos en kilogramos a libras, se tiene que multiplicar cada valor de la variable peso por la constante *k* =2.2.

   $$\sum_{i=1}^{5} k \times x_i = 2.2 \times \sum_{i=1}^{5} x_i = 2.2 \times (60.5 + 55 + 72.8 + 80.9 + 55) = 713.24 \; libras$$

3. $\sum_{i=1}^{n} (x_i - k) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} k = \sum_{i=1}^{n} x_i - n \times k$

   Ejercicio: Dada la constante *k* =3, obtenga:

   $$\sum_{i=1}^{5} (x_i - k) = 324.5 - (5 \times 3) = 309.2 \; kg$$

4. $$\sum_{i=1}^{n}(x_i - k)^2 = \sum_{i=1}^{n}(x_i^{\,2} - 2\,x_i \times k + k^2) = \sum_{i=1}^{n} x_i^{\,2} - 2k\sum_{i=1}^{n} x_i + n \times k^2$$

5. Otra operación frecuentemente utilizada envuelve la sumatoria del producto de dos variables, esto es, suponiendo que se tienen dos variables X y Y, cada una teniendo *n* observaciones, entonces:

$$\sum_{i=1}^{n} x_i\; y_i = x_1\; y_1 + x_2\; y_2 + \ldots + x_n\; y_n$$

Nota: $$\sum_{i=1}^{n} x_i\; y_i \neq \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i$$

Ejemplo 2: Sean X = variable peso expresado en kilogramos y Y = estatura en metros de *n*=5 alumnos:

i	1	2	3	4	5
x_i	60.5	55.0	72.8	80.9	55.0
y_i	1.60	1.69	1.85	1.58	1.76

Obtenga: $\sum_{i=1}^{n} x_i\; y_i$ $$\sum_{i=1}^{n} x_i\; y_i = (60.5 \times 1.60) + (55 \times 1.69) + \ldots + (60 \times 1.76) = 531.92$$

6. La sumatoria de los valores de dos variables es igual al resultado de la sumatoria de los valores de cada variable sumados uno al otro.

$$\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i} + \mathrm{y_i}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{x_i} + \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{y_i}$$

7. La sumatoria de una diferencia entre los valores de dos variables es igual a la diferencia entre los valores de las sumatorias de cada variable.

$$\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i} - \mathrm{y_i}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{x_i} - \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{y_i}$$

Ejemplo 3: Los datos de la siguiente tabla se refieren a los pesos en kilogramos al momento de nacer (Y) y al momento del destete (X) de *n*=6 becerros de la raza Nelore:

x_i	y_i
48.4	25.3
49.7	26.9
49.2	26.5
50	27.4
50.6	27.9
48.7	25.8

$$\sum_{i=1}^{n} x_i = 296.6$$

$$\sum_{i=1}^{n} y_i = 159.8$$

Calcular: $$\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i} - \mathrm{y_i}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{x_i} - \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{y_i} = 136.8\ \mathrm{kg}$$

1.12.3 Sumatorias dobles y múltiples

Hasta el momento, se han estado estudiando la suma de variables con un único índice. En problemas prácticos, sin embargo, nos encontraremos con situaciones en las cuales la variable respuesta puede tener dos o más índices.

Ejemplo 4: Suponga que se tiene que describir la producción de leche de n=12 vacas según su edad i (i=1,2,3,4 edades) y a su raza j (j=1,2,3 razas),como en la Tabla 1.

Tabla 1. Producción de leche en kg de 12 vacas de la Hacienda "Río Bravo", según la edad y la raza.

Raza	Edad en años				Total por raza
	3 (i=1)	6 (i=2)	9 (i=3)	12 (i=4)	
Cebú (j=1)	2.5	4.0	3.2	1.5	11.2
Holstein (j=2)	4.7	8.2	7.0	5.8	25.7
Nelore (j=3)	2.9	2.9	2.5	1.2	9.9
Total por edad	10.1	15.5	12.7	8.5	46.8

y_{ij} = producción de leche en kilogramos observada en la vaca de edad i y raza j.

El valor y_{32} = 7.0 tiene el significado de que la producción de la vaca de 9 años de la raza Holstein fue de 7.0 kilogramos. El valor $y_{..}$ = 46.8, representa la producción total (o gran total). Genéricamente para los 12 animales se tiene la siguiente tabla:

Tabla 2 Generalización de la Tabla 1

Raza	Edad en años				Total por raza
	(i=1)	(i=2)	(i=3)	(i=4)	
(j=1)	$.y_{11}$	$.y_{21}$	$.y_{31}$	$.y_{41}$	$\sum_{i=1}^{n} y_{i1} = y_{.1}$
(j=2)	$.y_{12}$	$.y_{22}$	$.y_{32}$	$.y_{42}$	$\sum_{i=1}^{n} y_{i2} = y_{.2}$
(j=3)	y_{13}	$.y_{23}$	$.y_{33}$	$.y_{43}$	$\sum_{i=1}^{n} y_{i3} = y_{.3}$
Total por edad	$\sum_{j=1}^{n} y_{1j} = y_{1.}$	$\sum_{j=1}^{n} y_{2j} = y_{2.}$	$\sum_{j=1}^{n} y_{3j} = y_{3.}$	$\sum_{j=1}^{n} y_{4j} = y_{4.}$	$\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} = y_{..}$

A través de los totales marginales tenemos que:

a) Para la producción total de la raza Cebú:

$$\sum_{i=1}^{4} y_{i1} = y_{.1} = y_{11} + y_{21} + y_{31} + y_{41} = 11.2 \text{ kg}$$

b) Para la producción total de las vacas con 3 años de edad:

$$\sum_{j=1}^{3} y_{1j} = y_{1.} = y_{11} + y_{12} + y_{13} = 10.1\, kg$$

c) Para la producción total

$$\sum_{i=1}^{4}\sum_{j=1}^{3} y_{ij} = y_{..} = y_{11} + y_{12} + y_{13} + \ldots + y_{43} = 46.8\, kg$$

LISTA DE EJERCICIOS 1

1. Clasifique (marcando con una X en el espacio correspondiente) según su naturaleza (cualitativas, cuantitativas discretas, cuantitativas contínuas) las siguientes variables:

No.	Nombre de la variable	Cuantitativa		Cualitativa
		Discreta	Continua	
1	Perímetro del cráneo de una cabra.			
2	Equipo de fútbol de preferencia			
3	Opinión sobre el servicio de emergencia de un hospital			
4	Número de hijos en un núcleo familiar			
5	Tiempo (en días) en que una fruta madura			
6	Cantidad de restaurantes en la ciudad de Cobán			
7	Diámetro a la altura del pecho (cms) de un árbol			
8	Barrio en el que vive un estudiante.			
9	Temperatura diaria de la ciudad de Escuintla			
10	Volumen de madera de un bosque de *Pinus oocarpa*			
11	Peso seco de las hojas de un árbol de encino			
12	Número de árboles muertos en una hectárea de bosque.			
13	Edad (en años cumplidos) de un grupo de alumnos			
14	Presencia de enfermedades respiratorias en niños			
15	Toneladas de caña producidas por hectárea			
16	Ingreso per cápita en Guatemala			
17	Grado de ataque de una virosis vegetal			
18	Cantidad de colonias de microorganismos sobre un fruto			
19	Orientación del viento			
20	Contenido de proteína (alto, medio, bajo) en leche			
21	Cociente entre largo y ancho de vainas de arveja			
22	Relación materia seca de parte aérea y raíces de soya			
23	Longitud del foliolo de hojas de garbanzo			
24	Color del tegumento de semillas de manía			
25	Zonas de vida presentes en el departamento de Petén			
26	Distancia (km) de Coatepeque a sus aldeas			
27	Características organolépticas de la carne de carnero			
28	Nivel de ausentismo laboral			
29	Compresión de la madera (kg.cm^2)			
30	Capacidad de Intercambio Catiónico (en meq/100 grs)			
31	Grupo sanguíneo de un grupo de personas			
32	Medidas de aislamiento: lavado de manos, uso de guantes, uso de mascarilla			
33	Agentes causantes de quemaduras: físicos, químicos			
34	Número de centros de salud por municipio			
35	Servicios hospitalarios: pediatría, neurología, medicina general.			
36	Material de construcción de las paredes de las casas			
37	Cantidad de computadoras por hogar en Escuintla			
38	Número de acciones diarias vendidas en la Bolsa Agrícola Nacional de Valores de Guatemala.			

2. Determine en que escala se expresa habitualmente cada una de las siguientes variables:

No.	Nombre de la variable	Escala
1	Número de llamadas telefónicas realizadas en un día x	
2	Horario de visita (mañana, medio día, tarde) de los polinizadores en una plantación de manzana.	
3	Porcentaje de mortalidad de plantas en un vivero	
4	Riesgo de incendio en un día "x" en una plantación forestal	
5	Diversidad de especies arbóreas en la cuenca del río Itzapa	
6	Diámetro de las copas de árboles en una plantación de cítricos	
7	Número de palabra recordadas en una evaluación de inglés	
8	Km de carretera asfaltada en los municipios de Alta Verapaz	
9	Precipitación pluvial (en mm) registrada en marzo	
10	Tiempo (años antes y después de Cristo).	
11	Intensidad de los terremotos en Escala Richter.	
12	Coordenadas de latitud y longitud geográficas	
13	Densidad de población de los municipios de Huehuetenango	
14	Clasificación de los núcleos de población en rurales (menos de 2,000 habitantes), intermedios (de 2,001 a 10,000 habitantes) y urbanos (más de 10,000 habitantes)	
15	Preferencias de una muestra de población para pasar las vacaciones en ciertas ciudades (puntuación de 1 a 10).	
16	Nota (rango 0 a 10) de un estudiante en el curso de Matemáticas I	
17	Ángulo de un triángulo medido en grados.	
18	Intensidad de los vientos en Escala Beaufor	

3. Los datos siguientes se refieren a la altura y diámetro a la altura del pecho (DAP) de 30 árboles de *Pinus caribaea* Morelet muestreados en el proyecto de reforestación Saquichaj, Cobán, Alta Verapaz.

Árbol	DAP (cm)	Altura (m)	Árbol	DAP (cm)	Altura (m)
1	11.79	18.00	16	21.84	17.20
2	12.34	16.00	17	21.84	17.00
3	14.95	16.50	18	22.67	17.40
4	15.00	17.00	19	23.01	17.50
5	15.05	17.00	20	25.82	17.00
6	15.40	16.00	21	26.00	19.00
7	15.55	18.00	22	26.23	18.00
8	16.94	17.00	23	26.44	18.00
9	17.71	18.00	24	26.57	17.20
10	17.83	16.50	25	27.46	18.00
11	18.69	17.50	26	27.50	17.20
12	19.51	15.00	27	27.83	17.30
13	20.80	16.90	28	28.08	17.40
14	20.90	17.00	29	30.83	17.00
15	21.01	17.10	30	30.88	18.00

Tome las variables X: x_i, i = 1, . . . , 30 y Y: y_i, i = 1, . . . , 30, para describir respectivamente el DAP y la altura, para n = 30 (total de árboles), calcule:

a) El coeficiente de correlación lineal de Pearson

$$\hat{\rho} = \frac{\sum_{i=1}^{n} x_i\, y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\,\sqrt{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}}$$

b) Coeficiente angular de la regresión

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

c) Suma de cuadrados total

$$\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

d) Intercepto

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}\,\frac{\sum_{i=1}^{n} x_i}{n}$$

1. Conocidos los siguientes valores: $x_1= 2$; $x_2=5$; $x_3= 4$; $x_4=8$; $y_1=3$; $y_2=8$; $y_3= 10$; $y_4= 6$. Calcule:

$a)\ \sum_{i=1}^{4} x_i$ $\quad b)\ \sum_{i=1}^{4} y_i$ $\quad c)\ \left(\sum_{i=1}^{4} x_i\right)^2$ $\quad d)\ \sum_{i=1}^{4} x_i\, y_i$

$e)\ \sum_{i=1}^{4} x_i^{\,2}$ $\quad f)\ \sum_{i=1}^{4} x_i \sum_{i=1}^{4} y_i$ $\quad g)\ \sum_{i=1}^{4} x_i\, y_i^{\,2}$ $\quad h)\ \sum_{i=1}^{4} (x_i + y_i)\,(x_i - y_i)$

2. En la tabla cruzada (o de contingencia) siguiente se presenta el número de plantas de manía (*Arachis hypogaea*) según el grado de severidad de una plaga y práctica cultural del lote:

Prácticas culturales	Severidad			Total
	Baja	Moderada	Alta	
Con rotación	235	124	38	397
Buena preparación de la cama de siembra	169	84	18	271
Uso de agroquímicos	452	67	27	546
Total	856	275	83	

Calcule:

a) $\sum_{j=1}^{n} y_{1j} = y_{1.}$ $\quad$ b) $\sum_{i=1}^{n} y_{i2} = y_{.2}$ $\quad$ c) $\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} = y_{..}$

UNIDAD II
ESTADÍSTICA DESCRIPTIVA

La Estadística Descriptiva es el conjunto de técnicas estadísticas que se utilizan para describir, en forma numérica, gráfica y tabular el comportamiento de un conjunto de datos. Si la variable X es cuantitativa, se medirán su tendencia central y su dispersión, así como la simetría y la curtosis. Además las medidas de posición relativa (fractiles).

2.1 DATOS SIN AGRUPAR

2.1.1 MEDIDAS DE TENDENCIA CENTRAL

Una tendencia central es un valor que mide alrededor de cuál número están concentradas las observaciones de una variable cuantitativa (punto medio de una distribución de datos) Estas medidas se llaman también medidas de posición. Las medidas de tendencia central utilizadas son: media, mediana y moda.

a) Media aritmética

Casi siempre cuando nos referimos al "promedio" de algo, nos estamos refiriendo al valor de la media aritmética. Esto es verdadero en casos como la temperatura promedio en la ciudad de Totonicapán en el mes de octubre, la vida promedio de la batería de una computadora o de la producción promedio de maíz amarillo en una hectárea de tierra.

Definición: Sean x_1, x_2, x_3, . . . , x_n los n valores observados para una variable cuantitativa X. Entonces la media aritmética de la variable X para una muestra es dada por la siguiente expresión:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

En notación sumatoria el estimador de la media para datos sin agrupar se representa de la siguiente manera:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ejemplo 5

Suponga que se tienen las notas obtenidas por un grupo de 20 estudiantes en un examen de Estadística General y que sus valores (ordenados de menor a mayor) son:

15, 45, 47, 53, 58, 58, 60, 62, 67, 74, 75, 78, 80, 80, 81, 85, 85, 85, 90, 92.

Para obtener la media aritmética, se suman los valores de x_i y el resultado se divide entre el número de observaciones:

$$\bar{x} = \frac{1370}{20} = 68.5 \text{ puntos.}$$

Propiedades de la media aritmética

1. La suma de las desviaciones de un conjunto de datos con relación a su media es nula.

Ejemplo 6: Si se consideran los siguientes datos: 1, 2, 3. $\bar{x} = 2$, entonces: (1–2) + (2–2) + (3 –2) = 0.

Prueba: $$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\,\bar{x} = \sum_{i=1}^{n} x_i - n\,\frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

2. La suma de los cuadrados de las desviaciones de un conjunto datos con relación a una constante *k* es mínima cuando *k* es la media aritmética.

Ejemplo 7: Considérense los siguientes datos: 1, 2, 3. $\bar{x} = 2$.

K	Suma de los cuadrados de las desviaciones
1	5.00
1,5	2.75
2.0	**2.00**
2.5	2.75
3.0	5.00

Prueba: $S = \sum_{i=1}^{n}(x_i - \theta)^2$, entonces el valor de θ que minimiza es S es obtenido solucionando:

$\frac{\partial S}{\partial \theta} = 0$, así: $\frac{\partial S}{\partial \theta} = 2 \sum_{i=1}^{n}(x_i - \theta)\;(-1) = 0$, entonces: $\sum_{i=1}^{n}(x_i - \theta) = 0 =$

$$= \sum_{i=1}^{n} x_i - n\theta = 0 = \frac{\sum_{i=1}^{n} x_i}{n} = \theta \Rightarrow \theta = \bar{x}.$$

3. La media de un conjunto de datos a los cuales se les ha sumado o restado en cada elemento una constante *k*, es igual a la media original más o menos esta constante.

Ejemplo 8: Si se tienen los siguientes datos: 1, 2, 3, con $\bar{x} = 2$ y *k* =2; los nuevos datos al sumar *k* son: 3, 4, 5. Y la nueva media $\bar{x}^* = 4 = \bar{x} + k = 2 + 2$.

Prueba: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, haciendo $x_i^* = (x_i \pm k)$, se tiene que:

$$\bar{x}^* = \frac{\sum_{i=1}^{n} x_i^*}{n} = \frac{\sum_{i=1}^{n}(x_i \pm k)}{n} = \frac{\sum_{i=1}^{n} x_i \pm n\,k}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \pm k$$

4. Si se multiplican todos los datos por una constante k, la nueva media quedará multiplicada por *k*.

Ejemplo 9: considere los datos: 1, 2, 3, con $\bar{x} = 2$ y k =3; los nuevos datos al multiplicar k son: 3, 6, 9. Y la nueva media $\bar{x}^* = 6 = k\,\bar{x} = (3)\,(2)$.

Prueba: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, haciendo $x_i^* = (k \cdot x_i)$, se tiene que:

$$\bar{x}^* = \frac{\sum_{i=1}^{n} x_i^*}{n} = \frac{\sum_{i=1}^{n} (k\, x_i)}{n} = \frac{k \sum_{i=1}^{n} x_i}{n} = k\;\bar{x}$$

A pesar de que la media presenta excelentes propiedades, que la mantienen como una de las medidas más importantes en Estadística, en ciertos casos ella puede no ser el parámetro o estimador más adecuado para describir un conjunto de dados, esto puede ocurrir, entre otros casos, cuando:

a) Se presentan datos extremos, aberrantes o discrepantes (observaciones cuyos valores están distintamente abajo o arriba de la mayoría de las demás observaciones), que no son representativos del conjunto de datos bajo estudio.

b) La distribución de los datos es asimétrica, bimodal o multimodal.

Algo más sobre la media aritmética

En el caso de datos no negativos, la media no solo describe el punto medio de un conjunto de datos, sino que también establece un límite sobre su tamaño.

Si se multiplica por n ambos lados de la ecuación: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, se tiene que $\sum_{i=1}^{n} x_i = \bar{x} \times n$ por tanto ningún valor de x puede ser mayor que: $\bar{x} \times n$.

Ejemplo 10: Si el salario anual medio pagado a los tres ejecutivos principales de una empresa es de US$ 156,000 ¿es posible que uno de ellos reciba $500,000?

Solución: Dado que n = 3 y $\bar{x}$ = $156,000, se tiene que,

$$\sum_{i=1}^{n} x_i = 3 \times 156{,}000 = \$\,468{,}000$$

y es imposible que cualquiera de los ejecutivos gane más de esa cantidad.

Ejemplo 11: Si nueve alumnos aspirantes a entrar a la USAC promediaron 41 puntos en la parte verbal de una prueba, ¿cuántos de ellos como máximo pueden haber promediado 65 ó más?

Solución: Ya que n = 9 y $\bar{x}$ = 41, se tiene que $\sum_{i=1}^{n} x_i = 9 \times 41 = 369$, y puesto que 65 está contenido cinco veces en 369 [369= (5 × 35) + 44] se deriva que a lo sumo cinco de los nueve estudiantes aspirantes pueden haber promediado 65 o más.

Se puede generalizar el argumento de los ejemplos 10 y 11. Para cualquier conjunto de datos no negativos con media $\bar{x}$, la fracción de los datos que son mayores o iguales que el valor positivo k no puede

exceder la fracción $\bar{x}/k$. Utilizando este resultado, conocido como: teorema de *Markov*, responda las preguntas siguientes:

a) Si el peso adulto medio de una raza de perros es de 35 libras, ¿qué fracción como máximo puede tener un peso mayor de 40 libras?

R\ 35/40 = 0.875 (87.5 %)

b) Si los árboles de cítricos de un huerto tienen un diámetro medio de 16.0 cm. ¿qué fracción de los árboles como máximo, puede tener un diámetro de 24 cm. ó más?

R\ 16/24 = 2/3 (aproximadamente 67 %)

b) Media ponderada

Si los *n* valores observados para una variable cuantitativa están ponderados por los pesos p_1, p_2, p_3, . . . , p_n, entonces la media de variable X, también llamada media ponderada , es dada por:

$$\bar{x}_p = \frac{p_1 x_1 + p_2 x_2 + p_3 x_3 + \ldots + p_n x_n}{p_1 + p_2 + p_3 + \ldots + p_n}$$

En notación sumatoria la media ponderada es:

$$\bar{x}_p = \frac{\sum_{i=1}^{n} p_i x_i}{\sum_{i=1}^{n} p_i}$$

A cada dato x_i le damos la importancia representada por su respectivo p_i.

Ejemplo 12

Una compañía utiliza tres niveles de trabajo: no calificado, semicalificado y calificado, para la producción de dos de sus productos finales. La compañía desea saber el promedio del costo de trabajo por hora para cada uno de los productos. Los datos son presentados en la siguiente tabla:

Nivel de trabajo	Salario por hora (*x*)	Horas de trabajo invertido por unidad producida	
		Producto 1	Producto 2
No calificado	US$ 5.00	1	4
Semicalificado	US$ 7.00	2	3
Calificado	US$ 9.00	5	3

Respuestas: $$\bar{x}_p = \frac{(1\times 5)+(2\times 7)+(5\times 9)}{(1+2+5)} = 8$$

Producto 1: US$ 8.00
Producto 2: US$ 6.80

Ejemplo 13: **Aplicación de la media ponderada en Hidrología**

El método de los polígonos de *Thiessen* es utilizado para calcular la precipitación media en un área conocida. Este método es aplicable a zonas con una distribución irregular de estaciones y donde los accidentes topográficos no juegan un papel importante en la distribución de las lluvias. El método de *Thiessen* tratar de tener en cuenta la falta de uniformidad en la distribución de los pluviómetros mediante un factor de ponderación para cada uno de ellos. La precipitación media se obtiene de la siguiente manera:

a) Se dibuja la zona en estudio con la ubicación exacta de las estaciones que contiene y las circunvecinas.

b) Se unen estas estaciones con trazos rectos, tratando de formar triángulos cuyos lados sean de la mínima longitud posible.

c) Después que los triángulos hayan sido definidos, se trazan las mediatrices (líneas perpendiculares bisectrices a las líneas de la unión) de todos los lados, con lo que se formarán unos polígonos alrededor de cada estación, se mide el área de cada polígono, la cual se determina utilizando un planímetro u otro método.

d) La lluvia media es el promedio ponderado de acuerdo a las áreas de cada polígono. Y está dada por la siguiente ecuación:

$$Pm = \frac{\sum_{i=1}^{n} P_i A_i}{\sum_{i=1}^{n} A_i}$$

siendo:

Pm = Precipitación media.
P_i = Precipitación de cada estación contenida en un polígono.
A_i = Áreas parciales de cada polígono.

Calcule la precipitación media por el método de los polígonos de *Thiessen* para la siguiente cuenca hidrográfica.

Estación	Precipitación observada (mm) P_i	Área parcial (km^2) A_i	$P_i \times A_i$
A	800	14.0	11,200
B	1000	14.5	14,500
C	1470	17.7	26,019
E	1750	15.8	27,650
H	2480	7.0	17,360
I	2150	6.5	13,975
Sumatoria		75.5	110,704

$$\text{Pm} = \frac{110,704 \text{ mm} \times \text{km}^2}{75.5 \text{ km}^2} = 1,466 \text{ mm}$$

c) Media geométrica

En ocasiones se trabaja con cantidades que cambian en un cierto período, se necesita conocer una tasa promedio de cambio, como la tasa de crecimiento promedio en un período de varios años. En tales casos, la media aritmética simple resulta inapropiada, puesto que, lo que se necesita encontrar es la media geométrica, que se denota por el símbolo $\bar{x}_G$. Existen dos usos principales de la media geométrica:

1. Para promediar porcentajes, índices y cifras relativas y
2. Para determinar el incremento porcentual promedio en ventas, producción u otras actividades o series económicas de un periodo a otro.

Ejemplo 14

Considere por ejemplo, el crecimiento de una cuenta de ahorros. Supóngase que se depositan US$ 100 inicialmente y se dejan acumular intereses a diferentes tasas durante cinco años. El crecimiento se resume en la siguiente tabla.

Año	Tasa de interés (%)	Factor de crecimiento[/1]	Ahorros al final del año
1	7	1.07	US$ 107.00
2	8	1.08	115.56
3	10	1.10	127.12
4	12	1.12	142.37
5	18	1.18	168.00

[/1] El factor de crecimiento es dado por: 1 + (tasa de interés/100).

El factor de crecimiento es la cantidad por la cual se multiplica los ahorros al inicio de año para obtener el saldo final del mismo. Para encontrar el factor de crecimiento promedio correcto, se multiplican los factores de crecimiento de los cinco años y luego se obtiene la raíz quinta del producto. La fórmula para encontrar la media geométrica de una serie de número es:

$$\bar{x}_g = \left[\prod_{i=1}^{n} x_i\right]^{1/n} = \left[x_1\, x_2\, x_3 \ldots x_n\right]^{1/n}$$

Aplicando logaritmo natural, se obtiene:

$$\ln \bar{x}_g = \frac{1}{n}\left[\ln x_1 + \ln x_2 + x \ln_3 + \ldots + \ln x_n\right]$$

Cuyo antilogaritmo provee el valor de $\bar{x}_g$. Para el ejemplo propuesto, el valor del factor de crecimiento promedio se obtiene de la siguiente manera:

$$\bar{x}_g = \left[\prod_{i=1}^{n} x_i\right]^{1/n} = \left[1.07 \times 1.08 \times 1.10 \times 1.12 \times 1.18\right]^{1/5} = 1.1093$$

US$ 100 × 1.1093 × 1.1093 × 1.1093 × 1.1093 × 1.1093 = US$ 167.98 ≈ 168.

Ejemplo 15

Supóngase que las utilidades obtenidas por una compañía constructora en cuatro proyectos fueron de 3, 2, 4 y 6%, respectivamente. ¿Cuál es la media geométrica de las ganancias?
En este ejemplo *n* = 4 y así la media geométrica es determinada por:

$$\bar{x}_g = \sqrt[4]{(3 \times 2 \times 4 \times 6)} = 3.464101615$$

y así la media geométrica de las utilidades es el 3.46%. La media aritmética de los valores anteriores es 3.75%. Aunque el valor 6% no es muy grande, hace que la media aritmética se incline hacia valores elevados. La media geométrica no se ve tan afectada por valores extremos.

d) Media armónica

La media armónica es la recíproca de la media aritmética de los recíprocos del conjunto de datos. Dada una muestra de *n* elementos distintos, su media armónica se determina a través de:

$$\bar{x}_a = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Esta media es poco sensible a los valores grandes, pero muy sensible a los valores próximos a cero, ya que los recíprocos $1/x_i$ son muy altos.

La media armónica no tiene un uso muy extenso en el mundo científico. Suele utilizarse principalmente para calcular la media de velocidades, tiempos o en electrónica.

Ejemplo 16

Supóngase que una familia realiza un viaje en automóvil a un ciudad y cubre los primeros 100 km a 60 km/h, los siguientes 100 km a 70 km/h y los últimos 100 km a 80 km/h. Calcular, en esas condiciones, la velocidad media realizada.

$$\bar{x}_a = \frac{1}{\frac{1}{3}\left(\frac{1}{60}+\frac{1}{70}+\frac{1}{80}\right)} = 69.041$$

e) Media cuadrática

La media cuadrática, valor cuadrático medio o RMS (del inglés *root mean square*) es una medida estadística de la magnitud de una cantidad variable. Puede calcularse para una serie de valores discretos o para una función de variable continua. El nombre deriva del hecho de que es la raíz cuadrada de la media aritmética de los cuadrados de los valores. El cálculo de la media cuadrática consiste en elevar al cuadrado todas las observaciones (así los signos negativos desaparecen), en obtener después su media aritmética y en extraer, finalmente, la raíz cuadrada de dicha media para volver a la unidad de medida original. La desviación estándar es una media cuadrática. Es una ecuación muy utilizada para calcular el diámetro medio o cuadrático de los datos de diámetro (DAP) en poblaciones forestales.

$$\bar{x}_{RMS} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} = \sqrt{\frac{x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2}{n}}$$

Propiedad: hay una relación de orden de las medias obtenidas de una misma colección de valores H≤ G ≤ A ≤ C, donde H es la media armónica; G, la media geométrica; A, la media aritmética; C, la media cuadrática.

f) Mediana

La mediana Md_X de un conjunto de n observaciones x_1, x_2, x_3, . . . , x_n es el valor que se encuentra en el centro del conjunto de datos, cuando están dispuestos en orden creciente o decreciente. Es decir, que el 50% de los datos es mayor que la mediana y el 50% restante es menor. El valor de la mediana dependerá de sí el número n datos es par o impar:

a) Si n es impar, entonces la mediana se encuentra en la posición $\frac{n+1}{2}$, que es exactamente la posición que separa los datos en dos grupos de igual cantidad.

b) Si n es par, entonces la mediana estará entre la posición $\frac{n}{2}$ y la posición $\frac{n}{2}+1$, para que los datos se dividan en dos grupos de $n/2$ valores cada uno. Es usual entonces tomar la mediana como la media aritmética entre los datos $x_{\left(\frac{n}{2}\right)}$ y $x_{\left(\frac{n}{2}+1\right)}$ es decir:

$$Md_x = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Ejemplo 17

Considerando el ejemplo de las notas obtenidas por un grupo de 20 estudiantes universitarios:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$x_{(i)}$	15	45	47	53	58	58	60	62	67	74	75	78	80	80	81	85	85	85	90	92

Como el número de datos es 20, entonces la mediana será la media aritmética entre los datos que están en la posición 10 ($x_{(20/2)}$) y la posición 11($x_{((20/2)+1)}$). Estos datos son 74 y 75. Entonces la mediana es:

$$Md_x = \frac{74+75}{2} = 74.5 \text{ puntos}$$

***Nota*:**
La mediana por ser poco afectada por valores extremos o discrepantes (*outliers*) se acostumbra decir que es una medida más robusta que la media aritmética.

g) Moda

La moda es una medida de tendencia central diferente de la media, pero un tanto parecida a la mediana, pues en realidad no se calcula mediante algún proceso aritmético ordinario. La moda se define como el valor que más se repite en un conjunto de datos, o sea, el valor que ocurre con más frecuencia. Puede suceder que en ciertos casos no se presente un valor modal (entonces se trata de una distribución amodal), o se presente más de un valor modal (distribuciones multimodales). En Excel utilice la función **=moda.mult**, para los casos en que la distribución de datos tenga más de una moda.

Ejemplo 18

Un estudiante de EPS de la Facultad de Agronomía está realizando un diagnóstico de una comunidad del municipio de Tucurú, Alta Verapaz. Consulta los archivos del puesto de salud de la comunidad y anota el número de hijos por familia que ha utilizado el puesto. Los datos son los siguientes: 3, 4, 3, 4, 5, 1, 6, 3, 4, 5, 3, 4, 3, 3, 4, 3, 5, 5, 5, 5, 6, 11, 10, 2, 1, 2, 3, 1, 5 y 2.
Con esta información:

a) Organice una tabla de frecuencias.
b) Calcule la media, mediana y moda.

Solución: La tabla de frecuencias queda de la forma siguiente:

La moda es: 3 hijos por familia. Por ser el valor con mayor frecuencia.

No. hijos	Frecuencia	Frecuencia acumulada
1	3	3
2	3	6
3	8	14
4	5	**19**
5	7	26
6	2	28
10	1	29
11	1	30

Posición: (n/2 + (n/2 +1)/2) = 15 y 16, se ubica en esta categoría.

El valor de la mediana es de 4 hijos por familia.

La media aritmética se obtiene así:

$$\bar{x} = \frac{(3\times 1)+(3\times 2)+(8\times 3)+(5\times 4)+(7\times 5)+(2\times 6)+(1\times 10)+(1\times 11)}{30} = \frac{121}{30} = 4.033$$, aproximadamente 4 hijos por familia.

2.1.2 MEDIDAS DE DISPERSIÓN

Una medida de dispersión o de variabilidad mide qué tan diferentes o distantes son las observaciones de una medida de tendencia central (generalmente la media aritmética)

a) Rango o amplitud

Es la diferencia entre el mayor y el menor valor observado en un conjunto de datos, obtiene así:
Rango = $x_{\text{máx}} - x_{\text{min.}}$ Presenta el inconveniente de solamente tomar en cuenta los valores extremos del conjunto de datos.

b) Varianza

Cada población tiene una varianza, que se simboliza con σ^2 (sigma cuadrada). Para calcular la varianza de una población se divide la suma de las distancias al cuadrado entre cada elemento de la población y la media, entre el número total de observaciones de dicha población. La varianza es dada por la siguiente expresión:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

Para el caso de muestras, la varianza es dada por:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

El cálculo de la varianza muestral por medio de la función anteriormente propuesta puede ser muy laborioso. Se presenta a continuación una forma operacional para su obtención, sin que sea necesario calcular la media explícitamente.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2\bar{x}\,x_i + \bar{x}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - 2\frac{\sum_{i=1}^{n}x_i}{n}\sum_{i=1}^{n}x_i + n\frac{\sum_{i=1}^{n}x_i}{n}\frac{\sum_{i=1}^{n}x_i}{n}\right] =$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right]$$

La varianza, se expresa en *unidades al cuadrado* de los datos originales. Estas unidades no son intuitivamente claras o fáciles de interpretar. Por esta razón se hace un cambio significativo en la varianza para calcular una medida útil de la desviación, una que no dé problemas con las unidades de medida y, en consecuencia, que sea menos confusa. Esta medida se conoce como *desviación estándar*.

c) Desviación estándar

La desviación estándar de un conjunto de datos se define como la raíz cuadrada del valor de la varianza.

Ejemplo 19

En la siguiente tabla se presentan los datos referentes a la duración (expresada en horas) de 15 focos.

180	190	190	205	210
210	220	250	250	265
280	310	330	350	370

Calcule la varianza y la desviación estándar muestral.

$$s^2 = \frac{1}{14}\left[\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{15}\right] = \frac{1}{14}\left[1021250 - \frac{(3810)^2}{15}\right] = 3822.1429\ (\text{horas})^2$$

$$s = \sqrt{3822.1429} = 61.8235\ \text{horas}$$

d) Coeficiente de variación

El coeficiente de variación mide la variabilidad porcentual de los datos respecto a su media:

$$C.V.(\%) = \frac{s}{\bar{x}} \times 100$$

El coeficiente de variación sirve para comparar la variabilidad de diferentes variables, y es particularmente útil cuando:

a) Los datos están en unidades diferentes. Suponga como ilustración, que se desea comparar la dispersión de las distribuciones de los pesos, en kg, y las alturas, en cm, de las mujeres de cierta población.

b) Los datos están en las mismas unidades, pero los valores son muy diferentes.

Ejemplo 20

La siguiente tabla proporciona la media y la varianza de los pesos y alturas de las mujeres de cierta población:

Atributo	Media	Varianza	Coeficiente de variación (%)
Altura (X)	$\bar{x} = 168$ cm	$s^2(X) = 900\ (\text{cm})^2$	17.86
Peso (Y)	$\bar{y} = 53$ kg	$s^2(Y) = 90\ (\text{kg})^2$	17.90

Con la información contenida en la anterior tabla, se observa que aunque la varianza de la altura sea, para este ejemplo, 10 veces mayor que la varianza de los pesos, los coeficientes de variación son prácticamente los mismos para las dos muestras.

2.1.3 MEDIDAS DE POSICIÓN RELATIVA: Percentiles

El percentil de orden 100p (P_{100p}) de un conjunto de valores dispuestos en orden creciente, es un valor tal que (100p)% de las observaciones son menores o iguales a él, y 100 (1-p)% son mayores o iguales a él ($0 < p < 1$).

El percentil de orden 50 (P_{50}) es igual a la mediana. Los percentiles de orden 25, 50 y 75 representados por Q_1, Q_2 y Q_3 son llamados cuartiles.

CÁLCULO DEL PERCENTIL DE ORDEN 100 (P_{100p}) PARA DATOS NO AGRUPADOS

np entero: $P_{(100p)} = \frac{x_{[np]} + x_{[np+1]}}{2}$

np no entero: $P_{(100p)} = x_{[\text{int}(np)+1]}$

siendo int (.) la función que aproxima un número para abajo hasta el entero más próximo. Por ejemplo: int (1.9) = 1, int (1.5) =1, int (1.2) = 1.

Ejemplo 21

En la siguiente tabla, se presentan los valores correspondientes a la producción (en gramos) de hule seco por sangría, por planta de hule, en el área A de la Hacienda "Caballo Blanco", Génova Costa Cuca.

10.2	10.2	10.3	10.6	10.8	11.0	11.6	11.8	11.9	12.0
12.1	12.6	12.6	12.8	12.8	13.0	13.1	13.2	13.4	13.5
14.0	14.9	15.2	15.3	15.3	15.4	15.8	16.0	16.2	16.3
16.9	17.7	18.1	18.3	18.4	18.7	19.6	19.8	19.9	20.0
20.3	20.3	21.9	22.0	22.2	22.4	22.8	23.2	23.5	23.8
24.2	24.5	24.6	24.9	25.1	25.5	26.0	26.3	26.8	28.1

Con esta información establecer:

a) El valor de la producción que separa el 10% de las plantas de hule más productivas.

n = 60 datos
p = 0.90 (percentil 90)
np = 54

$P_{90} = \frac{x_{[54]} + x_{[55]}}{2}$ $P_{90} = \frac{24.9 + 25.1}{2} = 25 \text{ gr}$

b) Calcular el percentil de orden 2.5

n = 60 datos
p = 0.025
np = 1.5

$P_{2.5} = x_{[\text{int}(1.5)+1]} = x_{(2)}$ $P_{2.5} = 10.2 \text{ gr}$

c) Calcular el percentil de orden 97.5

n = 60 datos
p = 0.975
np = 58.5

$P_{97.5} = x_{[\text{int}(58.5)+1]} = x_{(59)}$ $P_{97.5} = 26.8 \text{ gr}$

d) Los resultados proporcionados por Infostat se presentan a continuación:

Medidas resumen

Estadísticas	hule seco (gr)
n	60.00
Media	17.66
D.E.	5.23
Var(n-1)	27.32
CV	29.59
Mín	10.20
Máx	28.10
Mediana	16.60
Q1	12.80
Q3	22.20
Percentil 90	24.90
Percentil 2.5	10.20
Percentil 97.5	26.80

CUARTILES Y DESVIACIÓN CUARTIL

Cuando se calcula la mediana de una serie de datos cuantitativos, éstos se ordenan y la mediana los divide en dos grupos con la misma cantidad de elementos: hay un grupo inferior y otro superior. Cada uno de esos grupos, que ya están ordenados, tiene a su vez una mediana. La mediana del grupo inferior se llama *primer cuartil*, denotado como Q_1 y la mediana del grupo superior se llama tercer *cuartil*, denotado como Q_3.

El *segundo cuartil* (Q_2) es la mediana original de la serie completa de datos. Véase que la función de los cuartiles es dividir los datos originales en cuatro grupos con la misma cantidad de datos cada uno. Así, habrá un primer grupo que contiene al 25% de los datos y que va desde el menor de los datos hasta Q_1. El segundo grupo contiene al 25% de los datos y va de Q_1 a la mediana. El tercer grupo contiene al 25% de los datos y va de la mediana a Q_3. Finalmente, el cuarto grupo contiene también al 25% de los datos y va de Q_3 hasta el mayor de los datos. La *desviación cuartílica* (Q) se calcula dividiendo el rango intercuartílico (RIQ) entre dos, así: $Q = \frac{Q_3 - Q_1}{2}$

Ejemplo 22

Considerando los datos del ejemplo 5, de las notas obtenidas por un grupo de 20 estudiantes universitarios del curso de Estadística General:

Q_1 (entre 5 y 6), Q_2 (entre 10 y 11), Q_3 (entre 15 y 16)

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$x_{(i)}$	15	45	47	53	58	58	60	62	67	74	75	78	80	80	81	85	85	85	90	92

Los cuartiles son: Q_1 = 58 puntos, *Md* = 74.5 puntos, Q_3 = 83 puntos.

Entonces la desviación cuartílica es: $Q = \frac{83-58}{2} = 12.5$ puntos

2.1.4 REPRESENTACIÓN GRÁFICA PARA DATOS SIN AGRUPAR

a) Diagrama de tallos y hojas (*stem and leaf*)

Un diagrama de tallos y hojas sirve para dar una idea de la distribución de los valores de una variable cuantitativa, pero sin perder los valores originales observados. Para su construcción, cabe decir, que no existe una regla fija, pero la idea básica es dividir cada observación en dos partes: la primera (tallo), es colocada a la izquierda de una línea vertical y estará constituida por los valores de los datos sin el último dígito. Por ejemplo, si los datos van de 1500 a 2000, entonces los tallos serán 150*, 151*, 152*, . . ., 250*. El asterisco significa que cada hoja estará dada por un solo dígito.

Las hojas, que constituyen la segunda parte del diagrama, son colocadas a la derecha de la línea vertical que las separa del tallo, y serán el último dígito para los datos que empiezan en el tallo correspondiente. Por ejemplo, a la derecha del tallo 23 se colocarían datos como 230, 231, 231, 232, 233, 235, y las hojas: 0 1 1 2 3 5.

Procedimiento para la construcción del diagrama de tallos y hojas

1. Definir la unidad de medida que dividirá cada valor en dos partes: tallo y hojas.

2. Escribir los tallos en orden creciente de magnitud, verticalmente, y pasar una línea vertical a la derecha de ellos.

3. Asociar cada tallo a su respectiva hoja.

4. Ordenar en cada tallo las hojas en orden creciente de izquierda (límite con la línea vertical) a derecha.

Ejemplo 23

En la siguiente tabla, se presentan los pesos (expresados en kilogramos) de 32 alumnos de la carrera de Ingeniería Forestal de la Escuela Superior de Agricultura "Luiz de Queiroz" (Piracicaba, SP, Brasil)

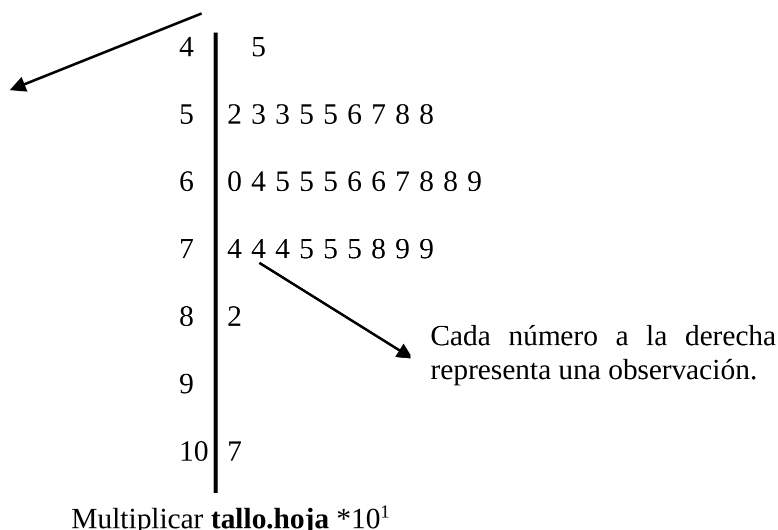
45	52	53	56	57	58	60	65	65
66	75	53	55	55	58	64	65	66
67	68	68	69	74	74	74	75	75
78	79	79	82	107				

Construya un diagrama de tallos y hojas para representar la distribución de este conjunto de datos.

La línea vertical divide los valores de las observaciones en una determinada unidad. En este diagrama, el lado izquierdo de la línea representa decenas de kilogramos (10 kg):

4 | 5 = 45 kg

10 | 7 = 107 kg



Multiplicar **tallo.hoja** $*10^1$

Por ejemplo: $4.5 * 10^1 = 45$, $5.2 * 10^1 = 52$, $5.3 * 10^1 = 53$, etc.

b) Diagrama de caja de dispersión (*box plot*)

Tanto la media como la desviación estándar, vistas anteriormente, pueden no ser medidas adecuadas para representar un conjunto de datos, pues, son muy afectadas por valores extremos, además, con esas dos medidas no se tiene idea de la forma de la distribución en cuanto a la simetría. Dada esta situación, Tukey (1977) sugirió la utilización de la mediana y los cuartiles Q_1 y Q_3 , cuya información puede ser traducida gráficamente en el llamado diagrama de caja de dispersión (*box plot*, *box-and-whisker plot*), que es una importante herramienta para el estudio de la simetría de las distribuciones y la detección de valores discrepantes. Los gráficos de *box plot* también son útiles para detectar, descriptivamente, diferencias en los comportamientos de grupos de variables. Por ejemplo podemos analizar el comportamiento de la precipitación pluvial en el municipio de Santa Lucía Cotzumalguapa (Escuintla), analizando una serie histórica de 20 años. Y luego construyendo un *box plot* por cada mes del año.

Procedimiento para la construcción del diagrama de dispersión

1. Ordenar los datos y calcular la mediana, los cuartiles Q_1 y Q_3 , y el rango intercuartílico (RIQ).

2. Dibujar una escala vertical, desde el menor hasta el dato mayor.

3. Dibujar un rectángulo de altura dado por Q_1 a Q_3 , y con base relativamente más pequeña. Luego, trazar una línea horizontal en el rectángulo, en el punto que corresponde a la mediana.

4. Calcular:
   $\boldsymbol{a} = Q_1 - 1.5RIQ$,
   $\boldsymbol{b} = Q_3 + 1.5RIQ$. Estos son los límites máximos permitidos a datos "típicos".

5. Calcular los últimos datos "típicos": $\boldsymbol{x_a}$ el menor dato observado que sobrepase $\boldsymbol{a}$ y $\boldsymbol{x_b}$ el mayor dato observado que no sobrepase $\boldsymbol{b}$. Es decir, $\boldsymbol{x_a}$ y $\boldsymbol{x_b}$ son los últimos datos observados considerados como "típicos".

6. Trazar una línea recta desde el rectángulo hasta $\boldsymbol{x_a}$ y otra desde el rectángulo hasta $\boldsymbol{x_b}$.

Ejemplo 24

Considerando los datos del ejemplo 5, de las notas obtenidas por un grupo de 20 estudiantes universitarios, y con los resultados del ejemplo 22, se tiene que:

Q_1 = 58 puntos, $Q_2 = Md$ = 74.5 puntos, Q_3 = 83 puntos y RIQ = 25 puntos.

Entonces:

$\boldsymbol{a}$ = Q_1 – 1.5*RIQ = 58 – 1.5 (25) = 20.5
$\boldsymbol{b}$ = Q_3 + 1.5*RIQ = 83 + 1.5 (25) = 120.5

Se espera que en el intervalo Q_1 – 1.5*RIQ y Q_3 + 1.5*RIQ se encuentre 99.3% de los datos.

Por lo tanto $\boldsymbol{x_a}$ = 45 (>20.5), es el último dato antes del límite permitido $\boldsymbol{a}$ = 20.5, y $\boldsymbol{x_b}$ = 92 (<120.5), por lo que en la parte superior todos los datos son "típicos". De acuerdo con los datos anteriores, la caja de dispersión que se obtiene es la mostrada en la Figura 1:

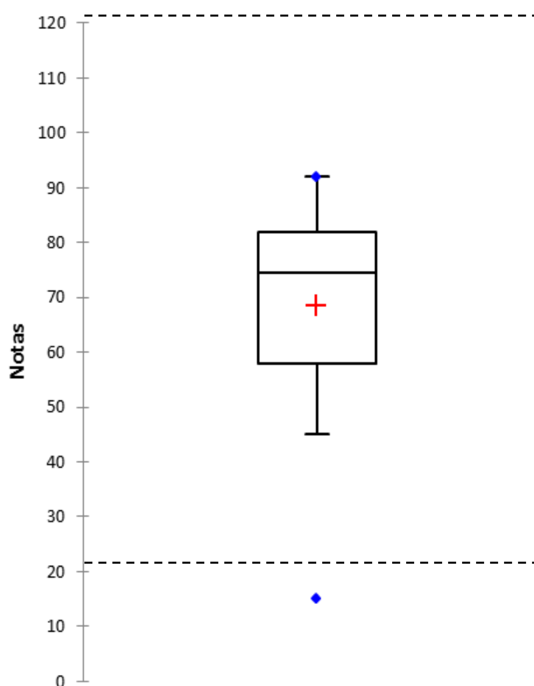


Figura 1 Diagrama de caja de dispersión para las notas obtenidas por 20 estudiantes universitarios.

Nota:

- El símbolo "+" representa a la media aritmética del conjunto de datos.
- En algunos programas computacionales estadísticos, los límites máximos permitidos a datos "típicos", se calculan así:

a) Límite para valores atípicos: Q_1 – 1.5*RIQ, Q_3 + 1.5*RIQ.
b) Límite para valores extremos: Q_1 – 3*RIQ, Q_3 + 3*RIQ.

A continuación se presentan algunas de las principales representación gráficas de los datos obtenidos en las investigaciones, es importante seleccionar aquellas que nos den la mejor representación visual de la información recabada, siendo esta, tarea del investigador. Recuerde que cada uno de estos gráficos posee normas para su construcción, las cuales debe de revisar cuidadosamente.

c) Diagrama de barras

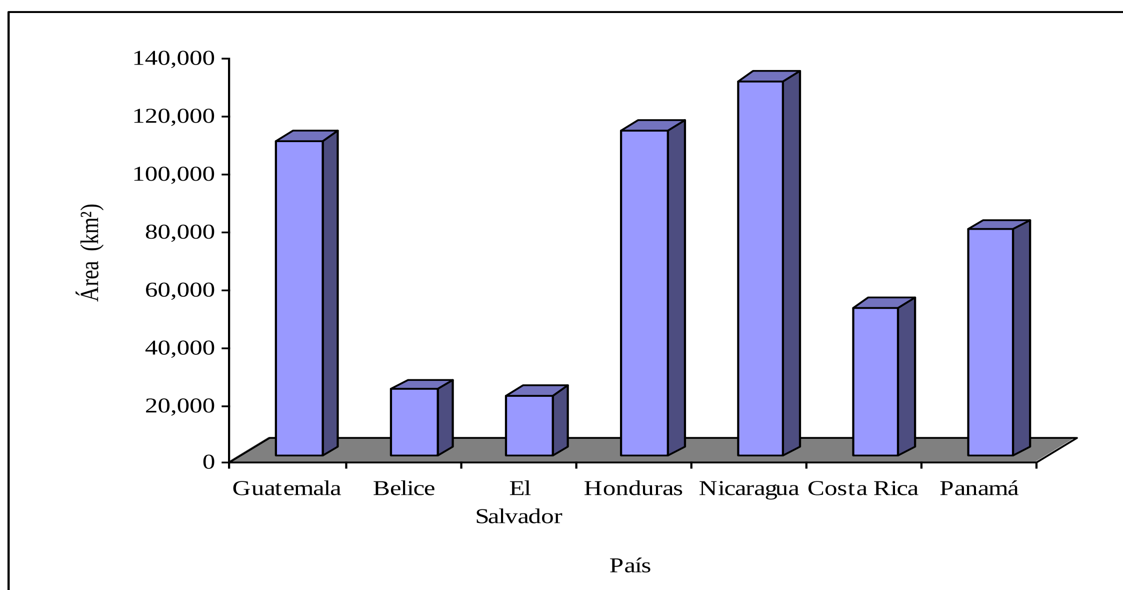


Figura 2. Extensión territorial de los países de Centro América.

d) Diagrama de barras compuestas

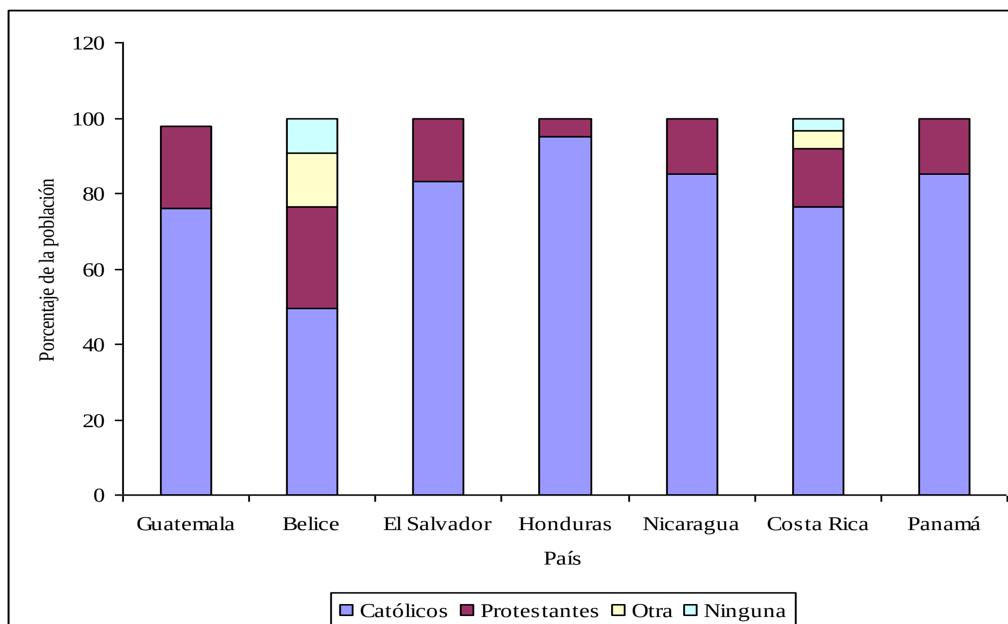


Figura 3. Principales religiones en los países de América Central, de acuerdo con la cantidad de adeptos.

e) Pastel, sectores o "*pie*"

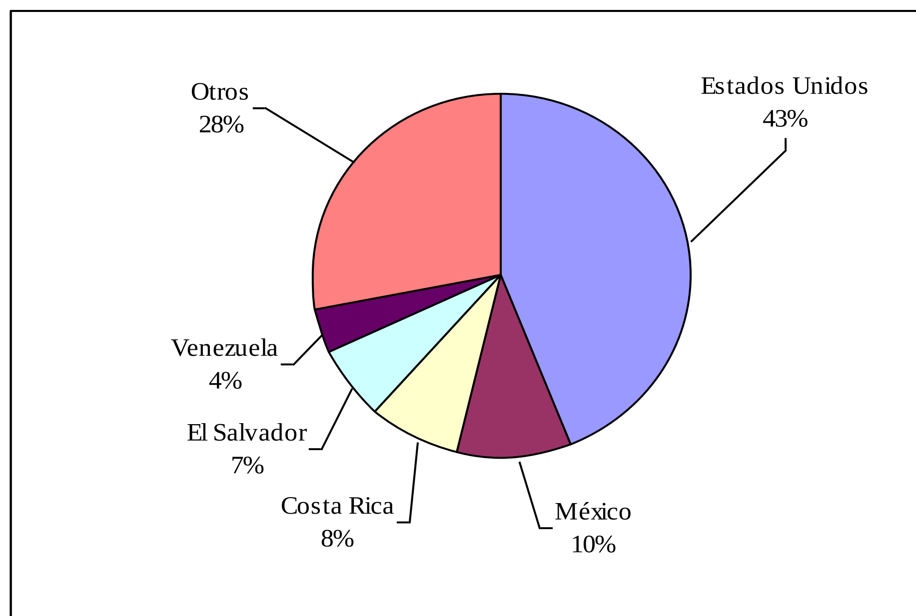


Figura 4. Principales países de los que Guatemala importa bienes.

f) Gráfico de barras y líneas

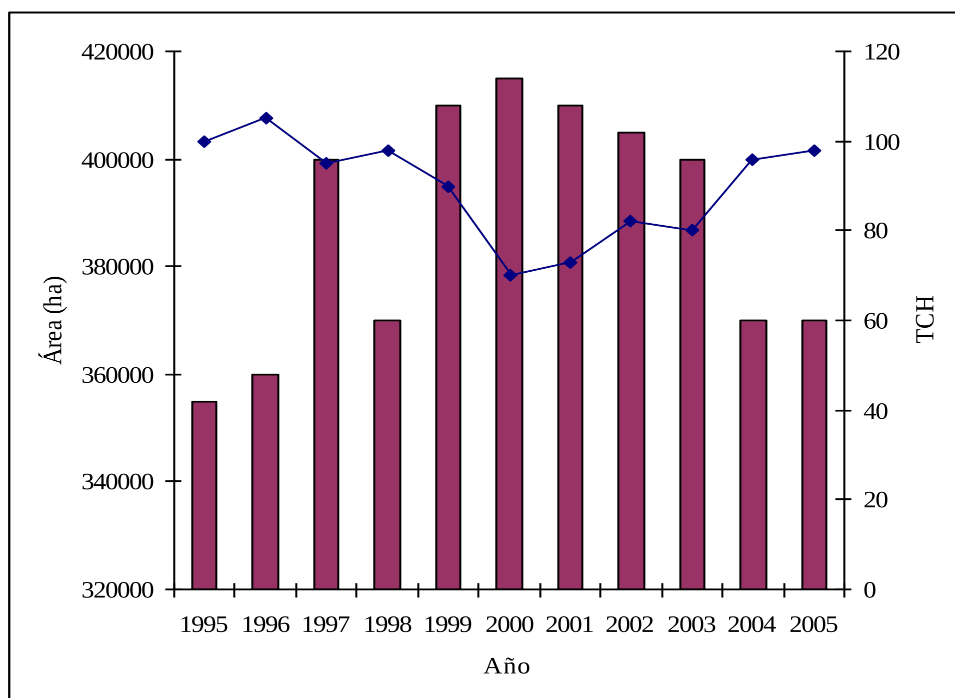


Figura 5. Extensión en hectáreas de caña cosechadas y rendimiento de toneladas de caña por hectárea obtenidas en los años 1995 a 2005 en Australia

g) Gráfico de líneas

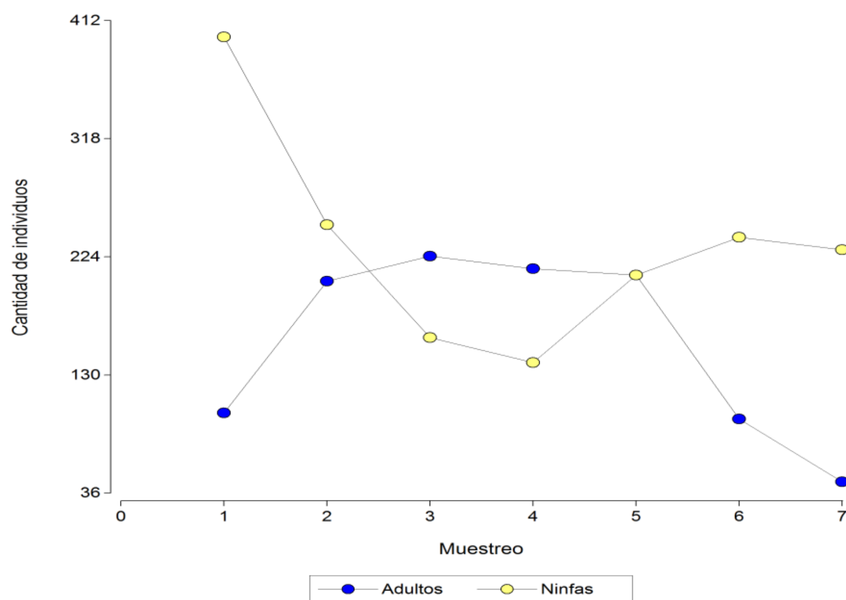


Figura 6. Comportamiento de la población de ninfas y de adultos de chinche salivosa (*Aeneolamia postica*) encontrados en 7 muestreos realizados en la finca "El Caobanal" del Ingenio "Concepción" (Escuintla), año 2006.

h) Gráfico de anillo o tipo dona

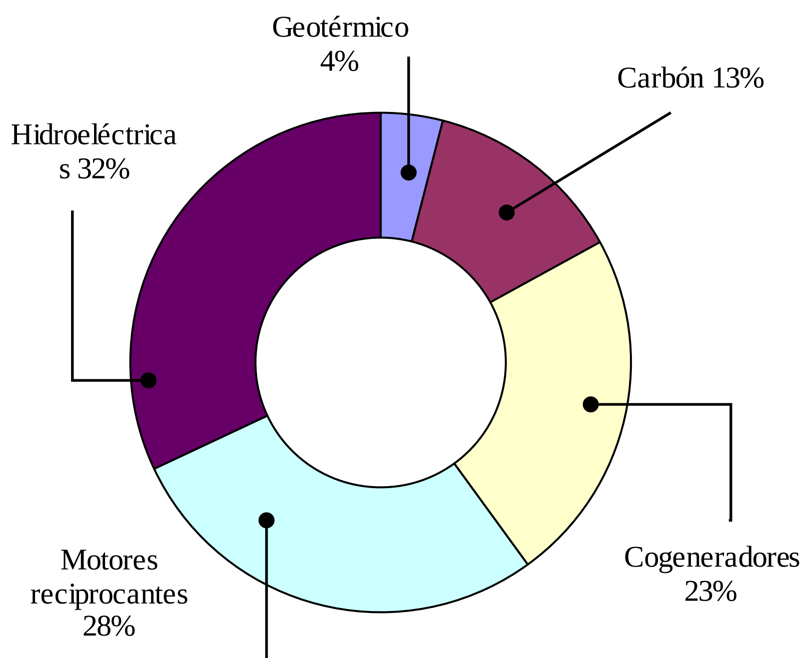
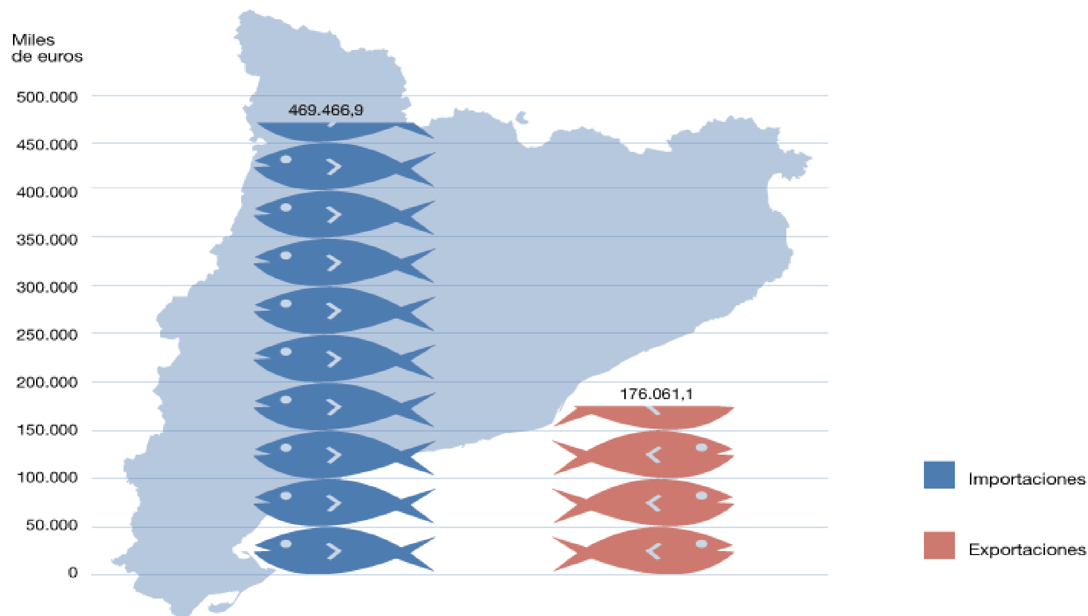
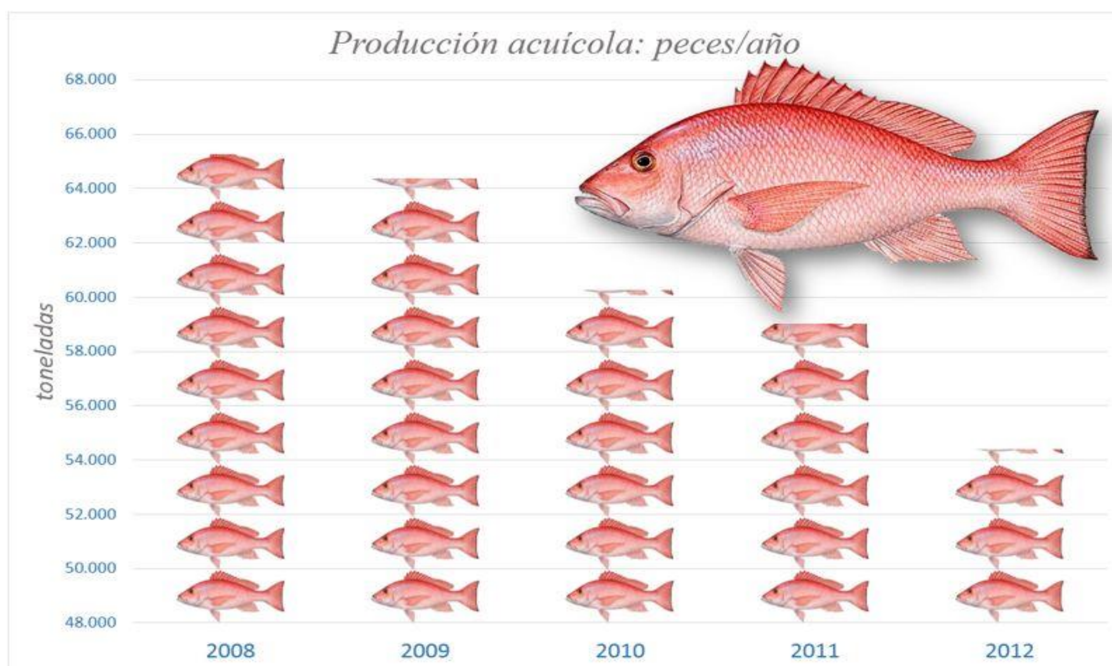


Figura 7. Composición de las fuentes energéticas en Guatemala, durante el mes de diciembre de 2007 (en porcentajes)

i) Pictogramas

Es un tipo de representación que se utiliza para variables cuantitativas, y que consiste en representar los datos con dibujos alusivos a la estadística estudiada. Los pictogramas son muy expresivos, pero poco precisos.



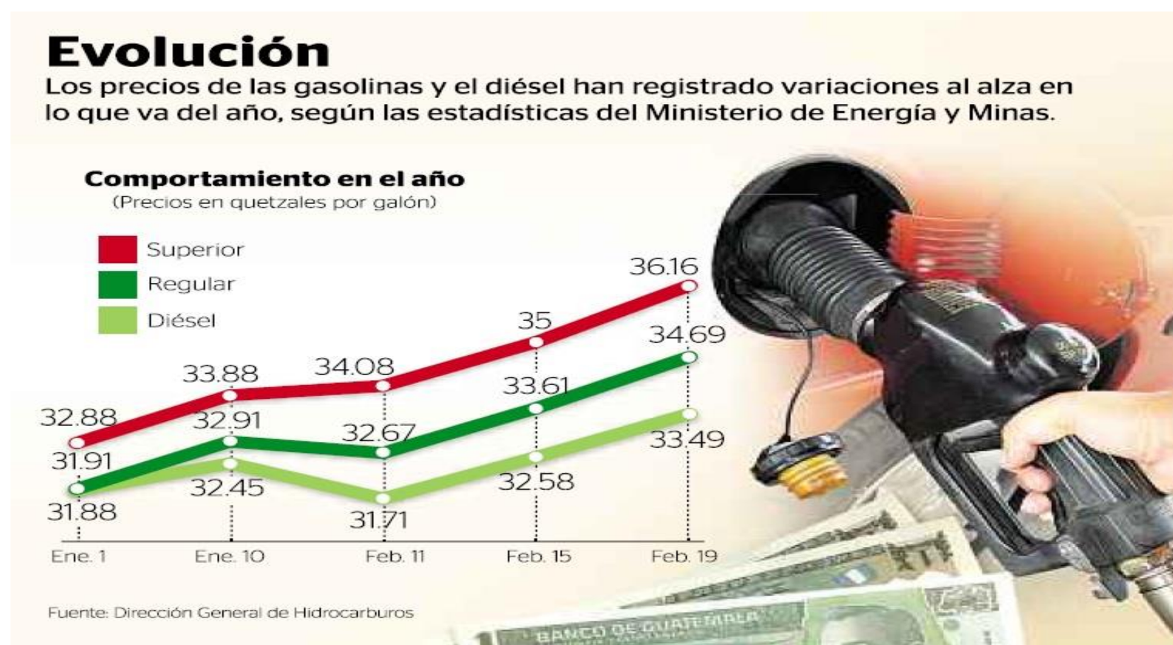


j) Infografías

Infografía define a una herramienta de comunicación que permite explicar visualmente procesos complejos en piezas gráficas de fácil comprensión, donde el texto da paso a elementos visuales explicativos. Las infografías son usadas para explicar de forma rápida y clara información o datos muy complejos (trayectorias, mapas, textos técnicos y material educativo)

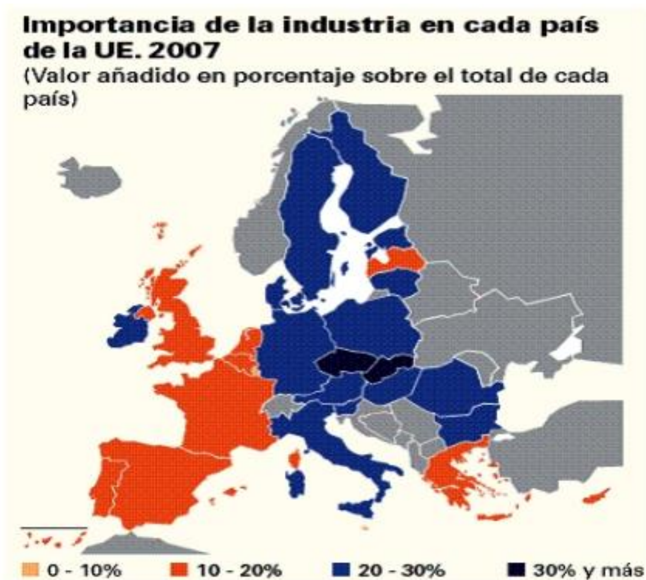
Esta herramienta gráfica es útil para desarrolladores de softwares, matemáticos e incluso estadistas que necesitan transmitir determinados procesos o hechos. A continuación se presentan algunos ejemplos de infografías, extraídos de publicaciones en periódicos del país.





k) Cartogramas

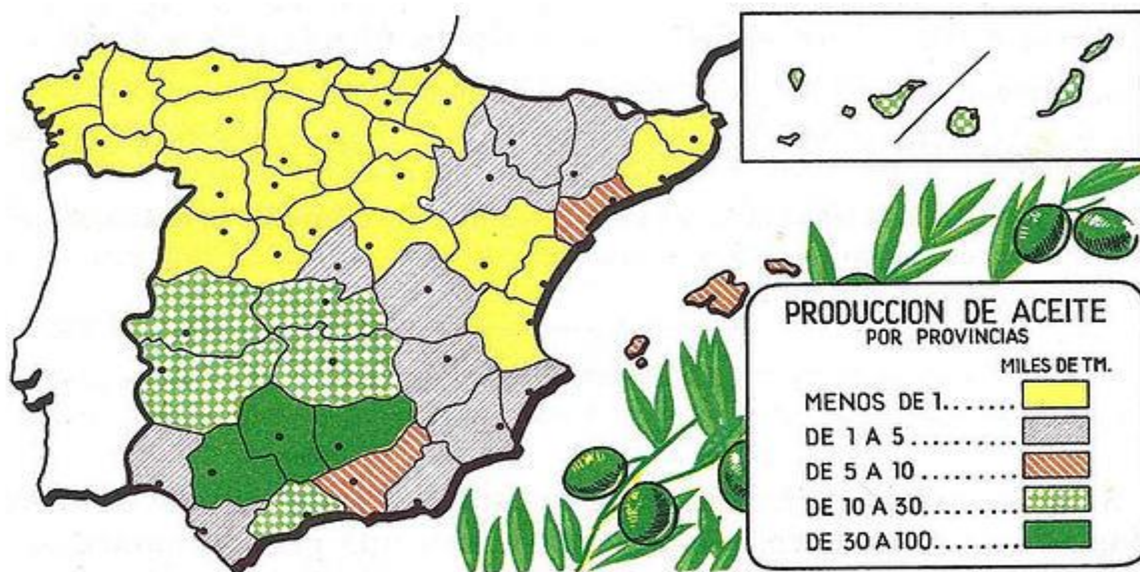
Un cartograma es un mapa en el que se presentan datos estadísticos por regiones, colocando el número o coloreando las distintas zonas en función del dato que representen. A continuación algunos ejemplos:



Fuente: Panorámica de la industria. INE



Fuente: España en cifras. INE



2.2 DATOS AGRUPADOS

Ejemplo 25

En la siguiente tabla se presentan los datos de diámetro a la altura del pecho (DAP) en centímetros, de *n*=280 árboles de la especie Santa María (*Callophyllum brasiliense*) en una extensión de 62.6 hectáreas de la finca "Agua Caliente".

25	27	29	32	34	36	39	44	48	59
25	27	29	32	34	36	39	44	48	59
25	27	29	32	34	37	39	44	49	61
25	27	29	32	34	37	39	44	49	63
25	27	29	33	34	37	39	44	49	63
25	27	29	33	34	37	40	44	49	66
25	27	29	33	34	37	40	44	49	68
25	27	30	33	34	37	40	45	49	68
26	28	30	33	34	37	40	45	49	80
26	28	30	33	35	37	40	45	50	80
26	28	30	33	35	37	40	45	50	
26	28	30	33	35	37	40	46	50	
26	28	30	33	35	37	41	46	50	
26	28	30	33	35	37	41	46	50	
26	28	30	33	35	37	41	46	51	
26	28	30	33	35	37	41	46	51	
26	28	30	33	35	38	41	46	51	
26	28	31	33	35	38	41	46	51	
26	28	31	33	35	38	41	47	52	
26	28	31	33	35	38	41	47	52	
26	28	31	33	35	38	42	47	52	
26	28	31	33	35	38	42	48	52	
26	29	31	33	35	38	42	48	52	
27	29	31	33	35	38	42	48	52	
27	29	32	33	36	38	42	48	52	
27	29	32	34	36	38	42	48	53	
27	29	32	34	36	38	43	48	53	
27	29	32	34	36	39	43	48	53	
27	29	32	34	36	39	44	48	54	
27	29	32	34	36	39	44	48	56	

Archivo: santa_maria.txt

El procedimiento consiste en:

1. Calcular el número de clases (NC) para la construcción de una tabla de frecuencias.

$$\begin{cases} NC = \sqrt{n} & NC = \sqrt{280} = 16.73 \approx 17 \text{, se utiliza cuando } n \leq 100 \\ NC = 2.5\ \sqrt[4]{n} & NC = 2.5\ \sqrt[4]{280} = 10.23 \approx 11\, clases \\ NC = 1 + 3.322 \log_{10} n & NC = 1 + 3.322 \log_{10}(280) = 9.13 \approx 10 \quad (\text{Ecuación de Sturges}) \end{cases}$$

(Esto es aproximadamente igual a 1 + $\log_2$ n)

La ecuación de Sturges fue propuesta por Herbert Sturger (1926). Para este caso el conjunto de datos se tiene que dividir en 10 clases.

Comentarios:

- Mendenhall (1990) cita que, normalmente es mejor utilizar de 5 a 20 clases.
- Otras ecuaciones para calcular el número de clases son:

$NC = 5\ \log_{10} n$ (cuando n >100), y el

Criterio de Scott (1979) basado en la normalidad de los datos. $NC = \frac{A \times n^{1/3}}{3.49 \times s}$, siendo:
A = amplitud total o rango, n = tamaño de la muestra y s = desviación estándar muestral.

2. Calcular la amplitud o rango = $x_{máx} - x_{min}$, en este caso: 80– 25 = 55 cm.

3. Calcular la amplitud o intervalo de clase (IC)

$$IC = \frac{\text{Rango}}{\text{No. Clases}} = \frac{55\, cm}{10 \text{ clases}} = 5.5\, cm / \text{clase}$$

(*) **Nota**: Es preferible que todas las clases tengan el mismo ancho o intervalo para poder hacer comparaciones uniformes de las frecuencias de clase.

4. Definir los límites de cada clase

Clases	Límite inferior	Límite superior
1	25.0	25.0 + 5.5 = 30.5
2	30.5	30.5 + 5.5 = 36.0
3	36.0	36.0 + 5.5 = 41.5
4	41.5	41.5 + 5.5 = 47.0
5	47.0	47.0 + 5.5 = 52.5
6	52.5	52.5 + 5.5 = 58.0
7	58.0	58.0 + 5.5 = 63.5
8	63.5	63.5 + 5.5 = 69.0
9	69.0	69.0 + 5.5 = 74.5
10	74.5	74.5 + 5.5 = 80.0

5. Construir la tabla de distribución de frecuencias

Clase (i)	LI	LS	Marca de clase (m_i)	Frecuencia (f_i)	Frecuencia acumulada (f_a) ↓	Frecuencia relativa (f_r)	Frecuencia relativa acumulada (f_{ra})	$m_i.f_i$	$m_i^2.f_i$
1	[25.0	30.5)	27.75	77	77	0.28	0.28	2136.75	59294.81
2	[30.5	36.0)	33.25	75	152	0.27	0.54	2493.75	82917.19
3	[36.0	41.5)	38.75	48	200	0.17	0.71	1860.00	72075.00
4	[41.5	47.0)	44.25	31	231	0.11	0.82	1371.75	60699.94
5	[47.0	52.5)	49.75	34	265	0.12	0.95	1691.50	84152.13
6	[52.5	58.0)	55.25	5	270	0.02	0.96	276.25	15262.81
7	[58.0	63.5)	60.75	5	275	0.02	0.98	303.75	18452.81
8	[63.5	69.0)	66.25	3	278	0.01	0.99	198.75	13167.19
9	[69.0	74.5)	71.75	0	278	0	0.99	0.00	0
10	[74.5	80.0]	77.25	2	280	0.01	1	154.50	11935.13
				280		1		10487.00	417957.00

Marca de clase: es el punto medio del intervalo de clase y se obtiene sumando los límites inferior y superior de la clase y luego dividiendo entre 2.

Frecuencia relativa (f_r) = frecuencia observada en la clase *i* / total de observaciones (*n*).

2.2.1 MEDIDAS DE TENDENCIA CENTRAL

a) Media aritmética

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i} \qquad \bar{x} = \frac{10487}{280} = 37.45 \text{ cm}$$

Siendo *k* = número de clases.

b) Moda

Cuando los datos están agrupados en una distribución de frecuencias, la moda es el punto del eje x, correspondiente a la ordenada máxima de la distribución. El proceso de cálculo de la moda en datos agrupados es el geométrico, a partir del histograma de frecuencias (método de Czuber). El método de Czuber está basado en la influencia que las clases adyacentes ejercen sobre la moda, la cual se desplaza en el sentido de la clase con mayor frecuencia.

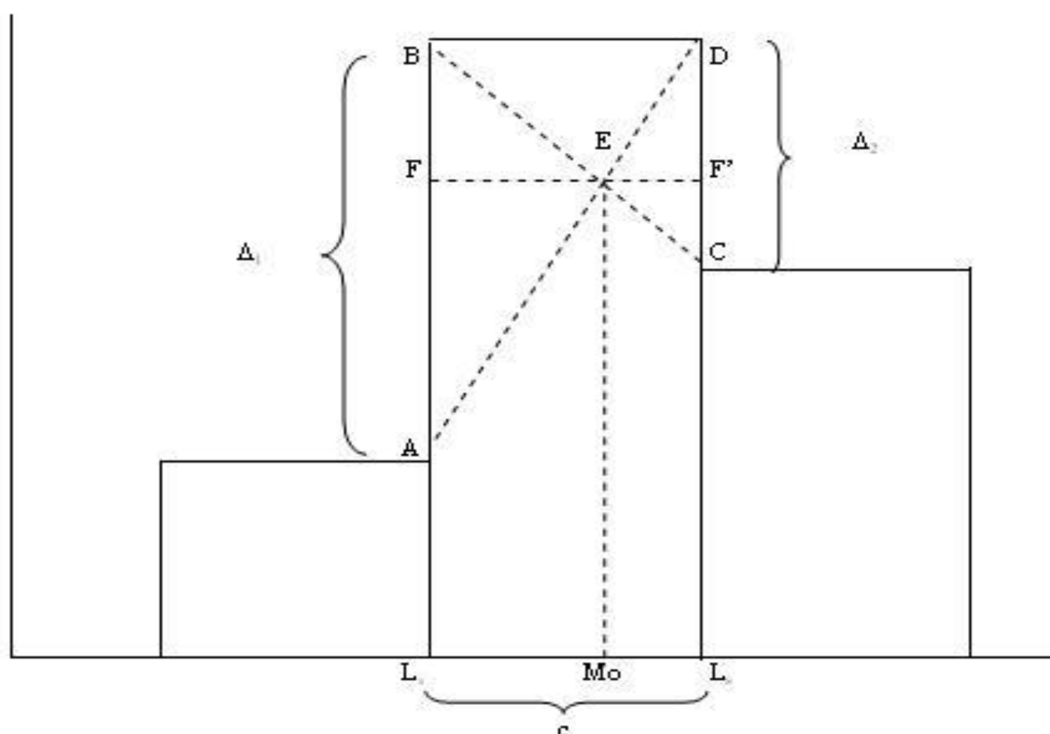


Figura 8 Esquema para la obtención de la moda por el método Czuber

En el histograma anterior se marca la clase modal, los vértices A, B, C y D. Se trazan las rectas $\overline{AD}$ y $\overline{BC}$. En el punto de intersección de estas rectas (E), se traza una perpendicular al eje de las clases, localizando el punto Mo, valor de la moda. El punto Mo divide el intervalo de la clase modal (c) en dos partes, cuyas longitudes son proporcionales a Δ_1 y Δ_2 Siendo Δ_1 la diferencia entre la frecuencia de la clase modal y la clase inmediata anterior, y Δ_2 la diferencia entre la frecuencia de la clase modal y la clase inmediata posterior.

Por E se traza la recta $\overline{FF'}$ paralela al eje de las clases, obteniéndose así, los segmentos $\overline{EF}$ y $\overline{EF'}$, que representan las alturas de los triángulos ABE y CDE. Siendo L_i el límite inferior de la clase modal, L*s* el límite superior y *x* la distancia entre L_i y la moda (Mo), se verifica en la Figura 2 que:

$$Mo = L_i + x$$

Siendo los triángulos ABE y CDE semejantes (revise el teorema de Thales de Mileto) se tiene que:

$$\frac{\overline{EF}}{\overline{AB}} = \frac{\overline{EF'}}{\overline{CD}} \Rightarrow \frac{\overline{EF}}{\overline{EF'}} = \frac{\overline{AB}}{\overline{CD}} \Rightarrow \frac{x}{c-x} = \frac{\Delta_1}{\Delta_2} \Rightarrow x\,\Delta_2 = \Delta_1\,(c-x)$$

$$\Rightarrow \quad x = \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c \;.$$

Como Mo = $L_i + x$, se tiene que: Mo = $L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$, en que:

L_i = límite inferior de la clase modal
Δ_1 = diferencia entre la frecuencia de la clase modal y la clase inmediata anterior
Δ_2 = diferencia entre la frecuencia de la clase modal y la clase inmediata posterior.
c = intervalo de clase.

Clase modal: es la clase que presenta la mayor frecuencia, es decir, es el valor más común.

Para el ejemplo 25, se tiene que la clase 1 es la clase modal, ya que tiene una frecuencia absoluta igual a 77 individuos. Siendo L_i = 25 cms.

$\Delta_1 = 77 - 0 = 77$ $\Delta_2 = 77 - 75 = 2$ $IC = 5.5$

Entonces, el valor de la moda es:

$$Mo = 25 + \left(\frac{77}{77+2} \right) \times 5.5 = 30.36 \ cm$$

c) Mediana

La mediana para datos agrupados se obtiene a partir de las frecuencias absolutas acumuladas, mediante interpolación lineal (teorema de Thales), tal como se ilustra la Figura 3.

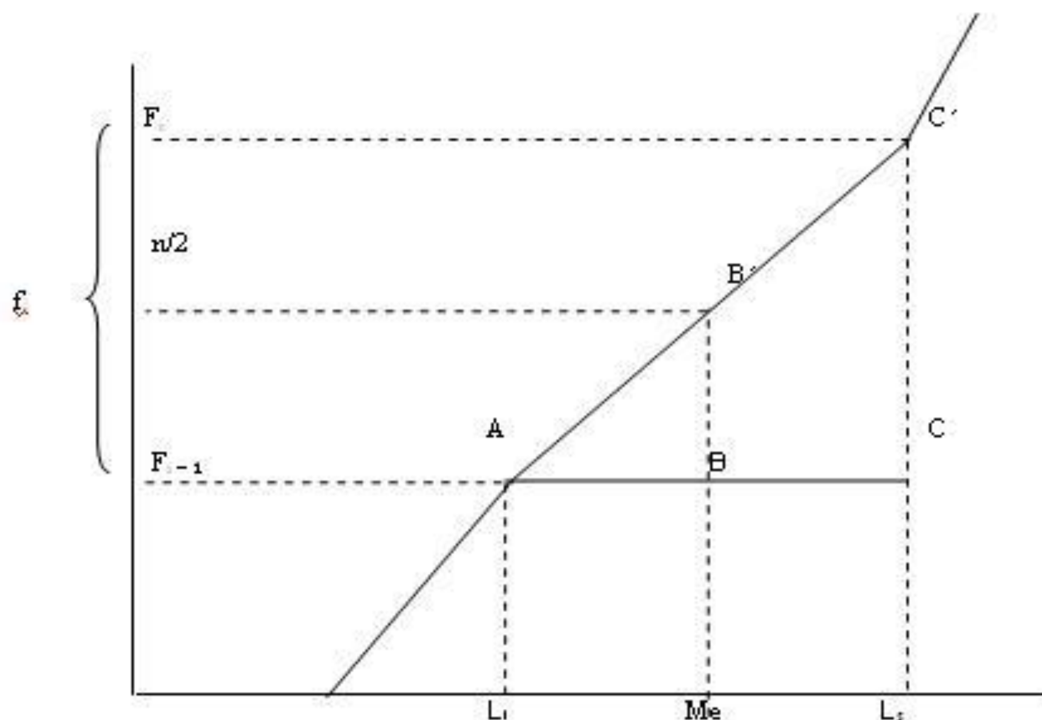


Figura 9 Esquema para la obtención de la mediana para datos agrupados

Siendo:

L_i = límite inferior de la clase donde se encuentra la mediana,
L_s = límite superior de la clase donde se encuentra la mediana,
F_i = frecuencia acumulada de la clase medianal,
F_{i-1} = frecuencia acumulada anterior a la clase medianal,
c = intervalo de clase.
f_i = frecuencia absoluta de la clase medianal.

Para los triángulos semejantes, ABB´ y ACC´ se tienen las siguientes relaciones:

$$\frac{\overline{CC'}}{\overline{AC}} = \frac{\overline{BB'}}{\overline{AB}} \quad \Rightarrow \quad \frac{\overline{CC'}}{\overline{BB'}} = \frac{\overline{AC}}{\overline{AB}} \quad = \quad \frac{f_i}{n/2 - F_{i-1}} = \frac{c}{Me - L_i}$$

$$\Rightarrow \quad Me - L_i = \frac{n/2 - F_{i-1}}{f_i} \times c \quad \Rightarrow \quad Me = L_i + \frac{n/2 - F_{i-1}}{f_i} \times c$$

Clase medianal: para identificar esta clase, en la columna correspondiente a las frecuencias acumuladas hacia abajo, se busca la clase que posea un valor de frecuencia acumulada igual o inmediatamente superior a *n* / 2. Para este ejemplo (280 / 2) = 140. Para el ejemplo 25, la clase 2 es la clase medianal, con L_i = 30.5, frecuencia acumulada de 152 y frecuencia absoluta = 75.

$$\text{Md} = 30.5 + \left(\frac{\frac{280}{2} - 77}{75} \right) \times 5.5 = 35.12 \text{ cm}$$

2.2.2 MEDIDAS DE DISPERSION

a) Varianza

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} m_i^2 f_i - \frac{\left(\sum_{i=1}^{k} m_i f_i\right)^2}{\sum_{i=1}^{k} f_i}\right] \qquad s^2 = \frac{1}{280-1}\left[417957 - \frac{(10487)^2}{280}\right] = 90.26 \text{ u}^2$$

b) Desviación estándar

$$s = \sqrt{90.26} = 9.5 \text{ cm}$$

c) Coeficiente de variación

$$\text{C.V.} = \frac{9.5}{37.45} \times 100 = 25.33\,\%$$

2.2.3 CÁLCULO DE PERCENTILES Y CUARTILES EN DATOS AGRUPADOS

Ejemplo 26

A continuación se presenta la tabla de distribución de frecuencias de los pesos (en kg.) de 310 pollos de engorde.

Límites de clase	Frecuencia absoluta	Frecuencia absoluta acumulada (↓)	Frecuencia absoluta acumulada (↑)
[0.8 – 1.0)	20	20	310
[1.0 – 1.2)	45	65	290
[1.2 – 1.4)	68	133	245
[1.4 – 1.6)	80	213	177
[1.6 – 1.8)	71	284	97
[1.8 – 2.0)	26	310	26
Total	310		

Para obtener el primer y el tercer cuartil se utilizan las siguientes ecuaciones:

$Q_1 = l_{Q1} + ((n/4 – faa))/f_{Q1}) \times IC$

$Q_3 = l_{Q3} + ((3n/4 – faa))/f_{Q3}) \times IC$

Siendo que: l_{Q1} y l_{Q3} son los límites inferiores de las clases que contienen al cuartil Q_1 y al cuartil Q_3, respectivamente; faa es la frecuencia acumulada anterior a la clase que contiene Q_1 o Q_3; f_{Q1} y f_{Q3} son las frecuencias absolutas simples de las clases que contienen los cuartiles e IC es el intervalo (o amplitud) de clase.

Para el ejemplo 26 tenemos que:

$$n/4 = 310/4 = 77.5 \quad y \quad 3n/4 = (3 \times 310)/4 = 930/4 = 232.5$$

De la misma forma como se obtuvo la mediana, para encontrar la clase que contiene el cuartil Q1, vemos en la columna de frecuencias absolutas acumuladas el valor superior a 77.5 (el más cercano) y vemos que es 133, por lo tanto, el límite inferior para esa clase l_{Q1}= 1.2. De igual manera, buscamos l_{Q3} y vemos que es: 1.6. Con esta información obtenemos los cuartiles Q_1 y Q_3.

$Q_1 = 1.2 + (77.5 – 65)/68) \times 0.2 = 1.237$ kg, y

$Q_3 = 1.6+ (232.5 – 213)/71) \times 0.2 = 1.655$ kg

Para obtener el i-ésimo percentil, i = 1, 2, . . . , 99, se utiliza la siguiente ecuación:

$P_i = l_{Pi} + ((i/100) \times n) – faa))/f_{Pi})* IC$

Así, por ejemplo, para obtener el segundo decil, tenemos:

$((20/100) \times 310) = 62$

$P_{20} = 1.0+ (62 – 20)/45) \times 0.2 = 1.187$ kg

2.2.4 DESCRIPCIÓN GRÁFICA

a) Histograma

El histograma es una secuencia de rectángulos colocados lado a lado, donde cada rectángulo tiene como base la amplitud (intervalo) de clase y como altura la frecuencia. Para el ejemplo 25 se tiene el siguiente histograma. Puede notar el comportamiento asimétrico de la distribución de los diámetros de los árboles.

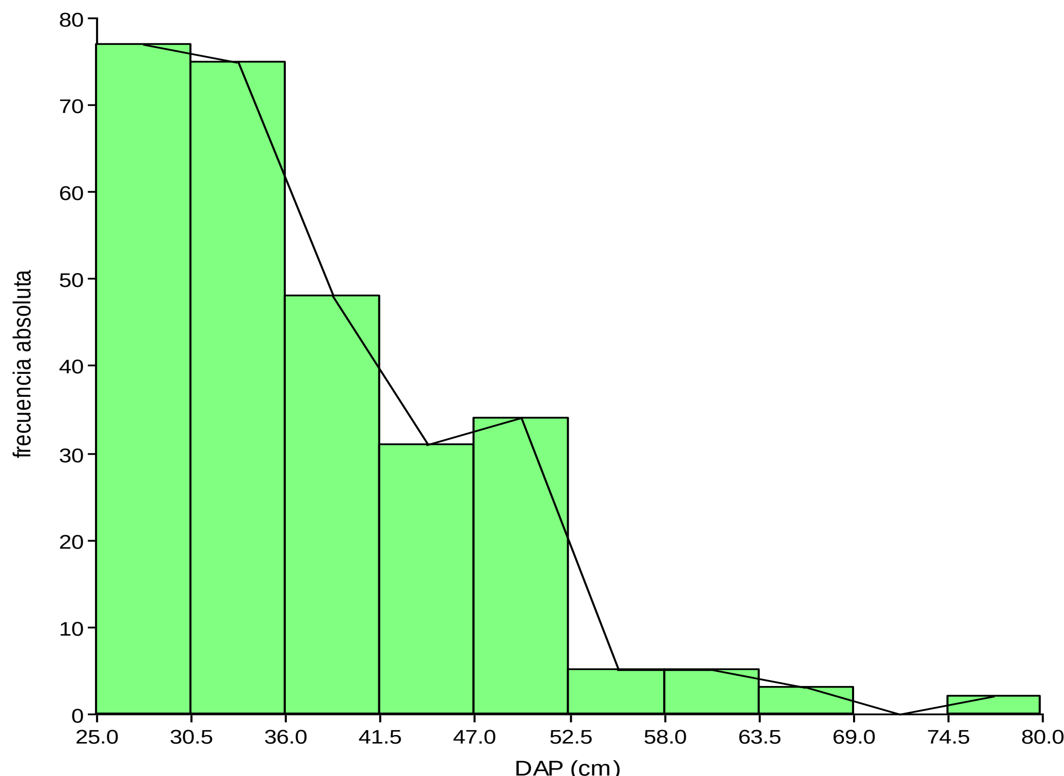


Figura 10 Distribución de los diámetros a la altura del pecho de 280 árboles de Santa María.

Los histogramas se utilizan frecuentemente para describir datos para los cuales no se ha realizado ningún tipo de supuesto. Si los datos se modelan como una muestra aleatoria proveniente de alguna distribución continua, el histograma de áreas puede ser considerado como un estimador de la función de densidad de probabilidad, pero tiene el inconveniente de no estar suavizado. Por lo que es necesario construir curvas de densidad utilizando por ejemplo, la estimación por núcleos de la función de densidad (kernel estimation), función núcleo gaussiana, función núcleo de Tukey biponderada, función núcleo de Epanechnikov, disponible en el lenguaje de programación R o en R Commander.

En la Figura 11 se presenta la función densidad para el conjunto de diámetros a la altura del pecho de los 280 árboles de Santa María.

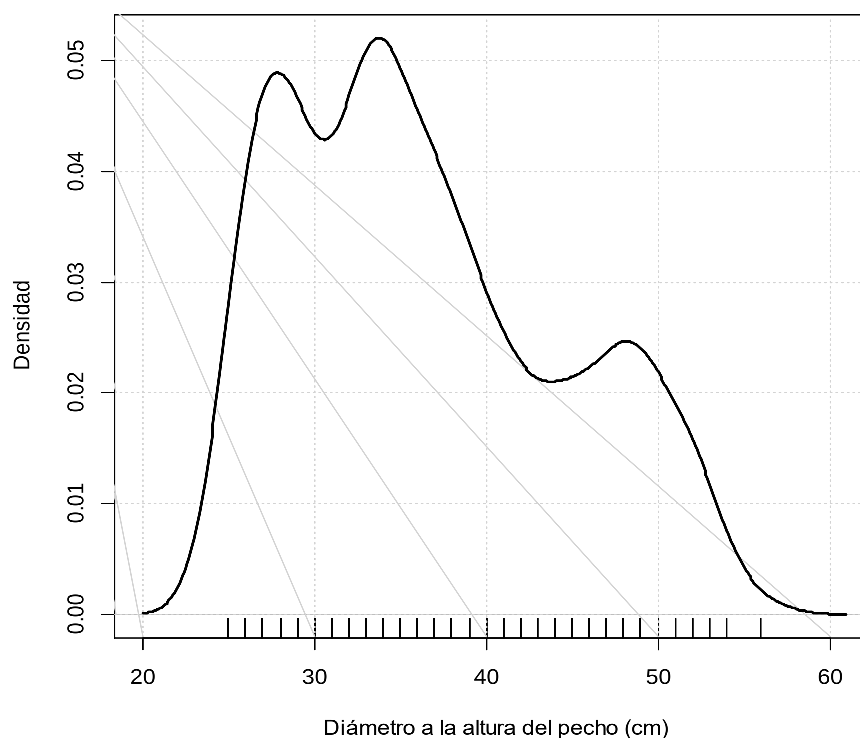


Figura 11 **Función de densidad para la distribución de los diámetros a la altura del pecho de 280 árboles de la especie Santa María**

b) Polígono de frecuencias

El polígono de frecuencias es un gráfico que se obtiene uniendo por una línea los puntos correspondientes a las frecuencias de las diversas clases, centradas en los respectivos puntos medios. Para obtener las intersecciones del polígono con el eje de la abscisa, se cría en cada extremo del histograma, una clase de frecuencias nula. En la Figura 10 junto con el histograma se presenta el polígono de frecuencias.

c) Ojiva de Galton

El gráfico de una distribución de frecuencias acumuladas o de frecuencias relativas acumuladas se conoce como Ojiva. Este tipo de gráfico permite ver cuántas o qué porcentaje de las observaciones están por debajo o por encima de ciertos valores. Las Figuras 12 y 13 muestran las ojivas de Galton de tipo "menor que" y "mayor que" para los datos del ejemplo 25.

La tabla de datos para poder construir ambas ojivas para el ejemplo 25 se presenta a continuación:

Clase (i)	Marca de clase (m_i)	Frecuencia (f_i)	Frecuencia acumulada (f_a) ↓	Frecuencia acumulada (f_a) ↑
0	22.25	0	0	
1	27.75	77	77	280
2	33.25	75	152	203
3	38.75	48	200	128
4	44.25	31	231	80
5	49.75	34	265	49
6	55.25	5	270	15
7	60.75	5	275	10
8	66.25	3	278	5
9	71.75	0	278	2
10	77.25	2	280	2
11	82.75	0		0

Clase "muerta" solo se construye para que cierre el polígono

Clase "muerta" solo se construye para que cierre el polígono.

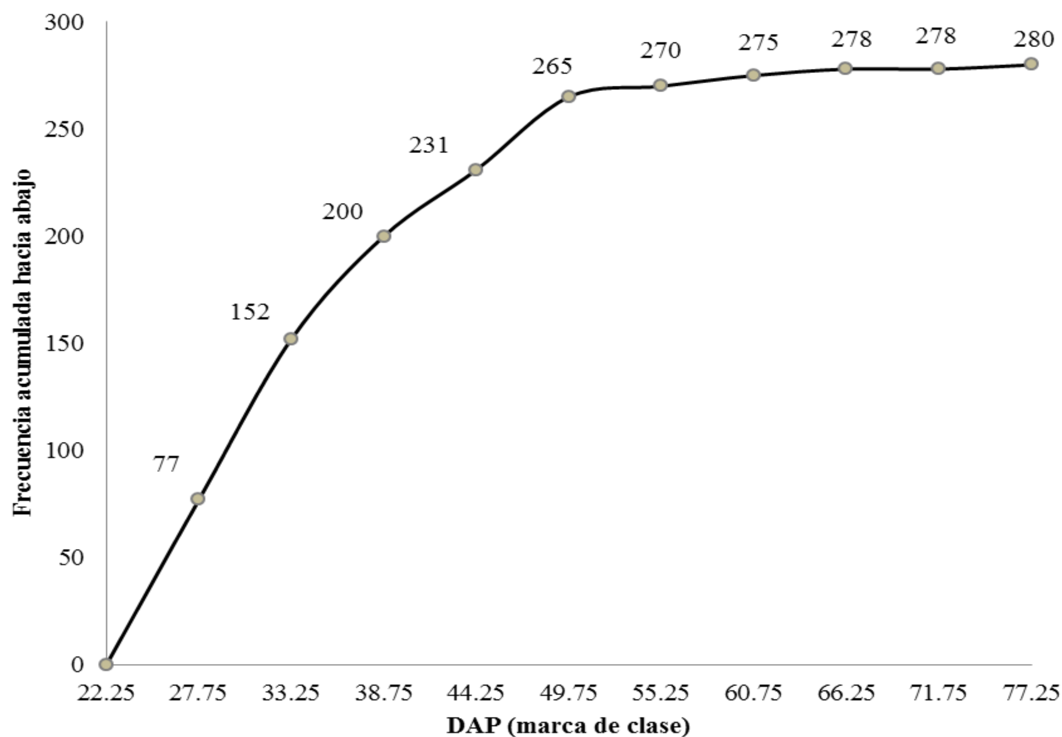


Figura 12 Ojiva de Galton de tipo "menor que" para la distribución de las frecuencias acumuladas de los diámetros de los árboles de Santa María.

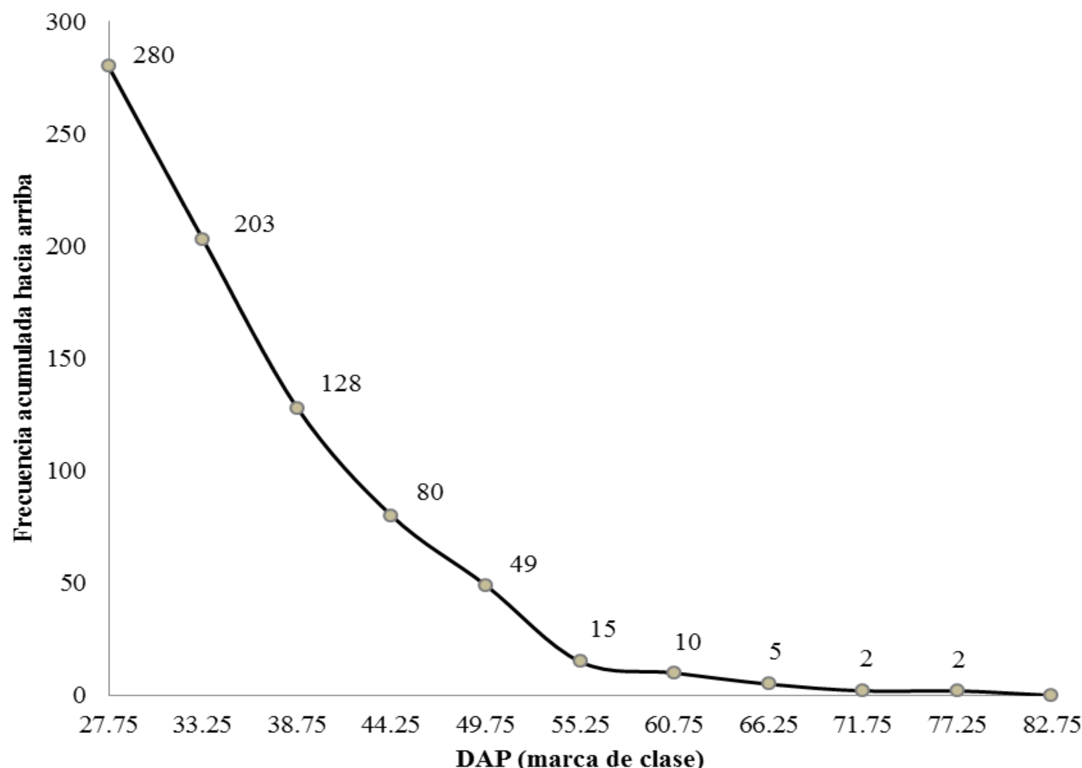


Figura 13 Ojiva de Galton de tipo "mayor que" para la distribución de las frecuencias acumuladas de los diámetros de los árboles de Santa María.

2.3 SIMETRÍA Y ASIMETRÍA

Las curvas de frecuencia que representan a un conjunto de datos pueden ser simétricas o sesgadas. Las curvas simétricas tienen una forma tal que una línea vertical que pase por el punto más alto de la curva dividirá el área de ésta en dos partes iguales. Cada parte es una imagen de espejo de la otra. Las curvas de frecuencia son sesgadas, cuando los valores no están igualmente distribuidos.

En resumen:

- La simetría de una distribución de frecuencias hace referencia al grado en que valores de la variable, equidistantes a un valor que se considere centro de la distribución, poseen frecuencias similares. Media y mediana coinciden en las distribuciones simétricas. Si sólo hay una moda (distribución unimodal), el valor de ésta también será igual a las dos anteriores.

- En distribuciones unimodales, el nivel de simetría se suele describir de acuerdo a tres grandes categorías: distribuciones simétricas, distribuciones asimétricas positivas (o sesgada a la derecha) y distribuciones asimétricas negativas (o sesgada a la izquierda).

- En caso de asimetría, los valores de la media, mediana y moda difieren. En concreto si la asimetría es positiva: media>mediana>moda. Si la asimetría es negativa: media<mediana<moda.

En las siguientes figuras se presentan curvas simétricas y sesgadas (asimétricas).

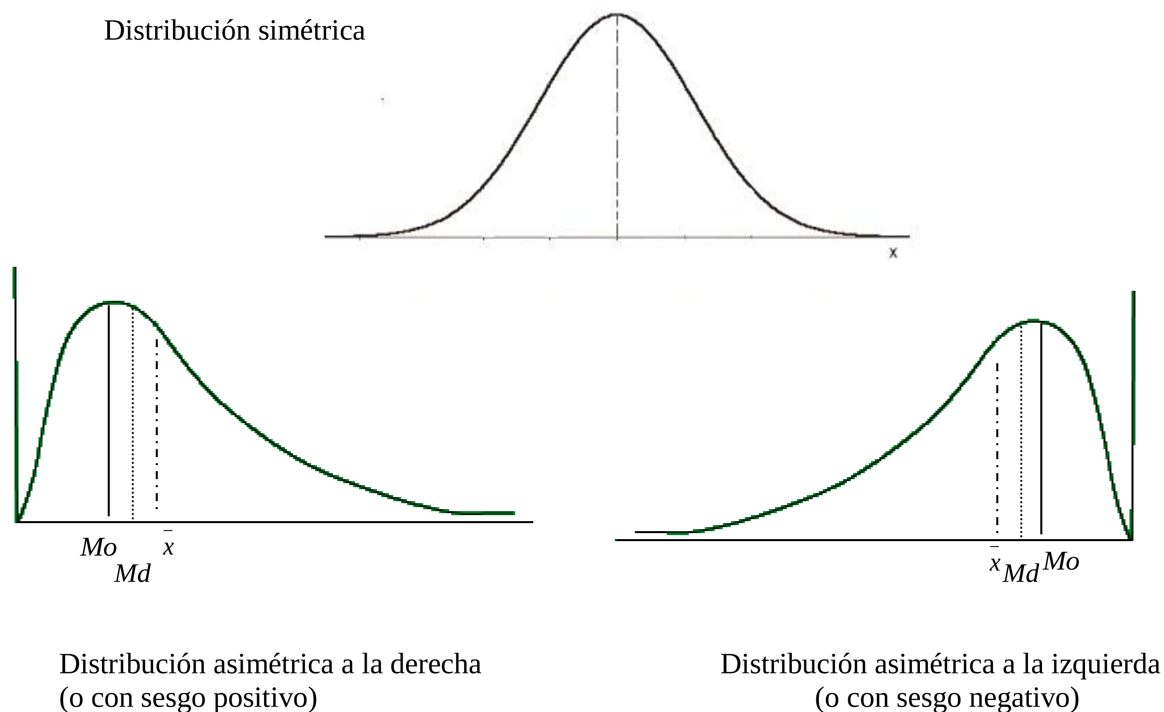


Figura 14 Tipos de asimetría o simetría en las distribuciones de datos.

¿Cómo medir el grado de asimetría de una distribución de datos?

Las medidas de asimetría tienen como finalidad el elaborar un indicador que permita establecer el grado de simetría (o asimetría) que presenta una distribución, sin necesidad de llevar a cabo su representación gráfica.

2.3.1 COEFICIENTE DE ASIMETRÍA DE PEARSON.

Karl Pearson propuso para distribuciones con forma de campana, unimodales y moderadamente asimétricas el coeficiente definido como As = ($\bar{x}$ – Mo) / s, donde Mo es la moda. Como en una distribución con forma de campana simétrica. $\bar{x}$ = Mo = Md, si la distribución es asimétrica positiva, $\bar{x}$ se desplaza a la derecha de la moda, y por tanto, $\bar{x}$ – Mo > 0. En el caso de distribución asimétrica negativa la media se sitúa por debajo de Mo, por lo que el valor $\bar{x}$ – Mo < 0.

La desviación estándar que aparece en el denominador no modifica el signo de la diferencia $\bar{x}$ – Mo y sirve para eliminar las unidades de medida de dicha diferencia. Así tendremos que si As = 0 la distribución es simétrica, si As > 0 la distribución es asimétrica positiva y si As < 0 la distribución es asimétrica negativa. También Pearson comprobó empíricamente para este tipo de distribuciones que se cumple: 3($\bar{x}$ – Md) ≈ $\bar{x}$ – Mo (la mediana siempre se sitúa entre la media y la moda en las distribuciones moderadamente asimétricas). Por esta razón, algunos autores utilizan como coeficiente de asimetría de Pearson la siguiente expresión:

$$As = \frac{3\left(\bar{x} - Me\right)}{s}$$

2.3.2 COEFICIENTE DE ASIMETRÍA DE FISHER.

$$g_1 = \frac{1}{n.s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Las fórmulas que utiliza el programa Excel para el cálculo de los coeficientes de asimetría y curtosis no son iguales a las utilizadas en esta sección. Por ejemplo, el coeficiente de asimetría se calcula mediante la expresión:

$$g_1 = \frac{n}{(n-1)(n-2).s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Esto produce una diferencia pequeña entre los cálculos anteriores y los realizados utilizando las funciones de Excel. Para el caso de datos agrupados, el coeficiente de asimetría de Fisher se calcula así:

$$g_1 = \frac{1}{n.s^3}\sum_{i=1}^{k}(m_i - \bar{x})^3.n_i$$

Si $g_1 = 0$ la distribución es simétrica, si $g_1 > 0$ la distribución es asimétrica positiva (a derecha), y si $g_1 < 0$ la distribución es asimétrica negativa (a izquierda). La distribución es asimétrica a derecha o positiva cuando la suma de las desviaciones positivas de sus valores respecto de la media es mayor que la suma de las desviaciones con signo negativo (la gráfica de la distribución tiene más densidad a la derecha de la media). En caso contrario, la distribución es asimétrica a la izquierda o negativa.

2.3.3 COEFICIENTE DE ASIMETRÍA DE BOWLEY (para variables ordinales)

Está basado en la posición de los cuartiles y la mediana, y viene dado por la expresión:

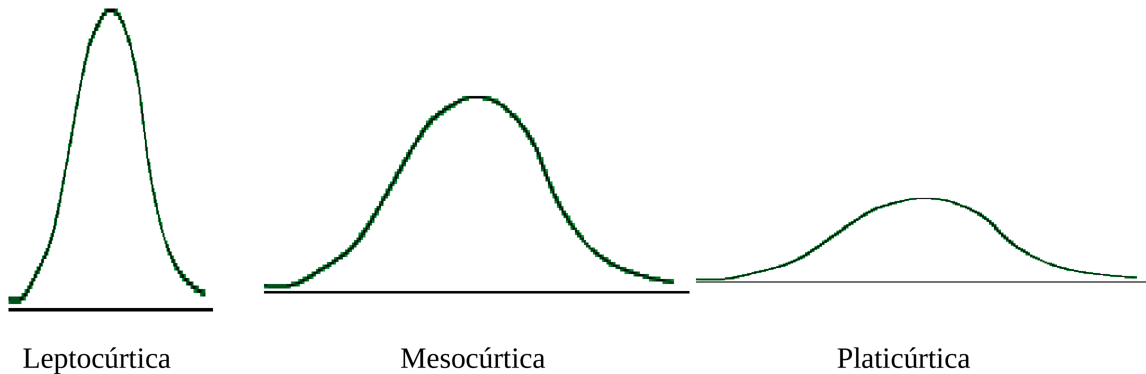
$$As(b) = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{Q_1 + Q_3 - 2Md}{Q_3 - Q_1}$$

2.4 CURTOSIS

Las medidas de curtosis estudian la distribución de frecuencias en la zona central de la misma. La mayor o menor concentración de frecuencias alrededor de la media y en la zona central de la distribución dará lugar a una distribución más o menos apuntada. Por esta razón a las medidas de curtosis se les llama también de apuntamiento o concentración central. Las medidas de curtosis se aplican a distribuciones campaniformes, es decir, unimodales simétricas o con ligera asimetría.

Para estudiar la curtosis de una distribución es necesario definir previamente una distribución tipo, que se toma como modelo de referencia. Esta distribución es la **Normal**, que corresponde a fenómenos muy corrientes en la naturaleza, y cuya representación gráfica es una campana de Gauss.

Tomando la normal como referencia, se dice que una distribución puede ser más apuntada que la normal (es decir, l**eptocurtica**) o menos apuntada (es decir, **platicúrtica**). A la distribución normal, desde el punto de vista de la curtosis, se le llama **mesocúrtica**. Con la curtosis se estudia la deformación, en sentido vertical, respecto a la normal de una distribución.



Leptocúrtica Mesocúrtica Platicúrtica

Figura 15 Tipos de curtosis en las distribuciones de datos.

¿Cómo medir la curtosis de una distribución de datos?

Coeficiente de curtosis*:* En la distribución normal se verifica que $m_4 = 3\sigma^4$, siendo m_4 el momento de orden 4 respecto a la media y σ la desviación estándar poblacional. Si se considera la expresión $g_2 = (m_4/\sigma^4) - 3$, su valor será cero para la distribución normal. Por ello, como coeficiente de apuntamiento o curtosis muestral se utiliza la expresión:

$$g_2 = \frac{1}{n.s^4}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Excel calcula el coeficiente de curtosis utilizando la siguiente expresión:

$$g_2 = \frac{n\,(n+1)}{(n-1)\,(n-2)\,(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3\,(n-1)^2}{(n-2)\,(n-3)}$$

..

Y para datos agrupados:

$$g_2 = \frac{1}{n.s^4}\sum_{i=1}^{k}(m_i - \bar{x})^4.n_i$$

Dependiendo del valor del coeficiente de curtosis, una distribución es:

a) **Mesocúrtica** (apuntamiento igual al de la normal) cuando $g_2 = 0$,

b) **Leptocúrtica** (apuntamiento mayor que el de la normal) sí $g_2 > 0$,

c) **Platicúrtica** (apuntamiento menor que el de la normal) sí $g_2 < 0$.

2.5 TEOREMA DE TCHEBYSHEV

El teorema establecido por el matemático ruso P.L. Tchebyshev (1821−1894), indica que, no importando que forma tenga la distribución de los datos, al menos 75% de los valores están comprendidos entre ± 2 desviaciones estándar a partir de la media de la distribución, y al menos 89% de los valores caen entre ± 3 desviaciones estándar a partir de la media. Pudiéndose medir aún con mayor precisión el porcentaje de observaciones que caen dentro de un alcance específico en curvas simétricas con forma de campana, se puede decir que:

1. Aproximadamente 68% de los valores de la población cae dentro de ± 1 desviación estándar a partir de la media.

2. Aproximadamente 95% de los valores estará dentro de ± 2 desviaciones estándar a partir de la media.

3. Aproximadamente 99% de los valores de la población estará en el intervalo que va desde tres desviaciones estándar por debajo de la media hasta tres desviaciones estándar por encima de la media.

Lo anterior se ilustra en la Figura 16:

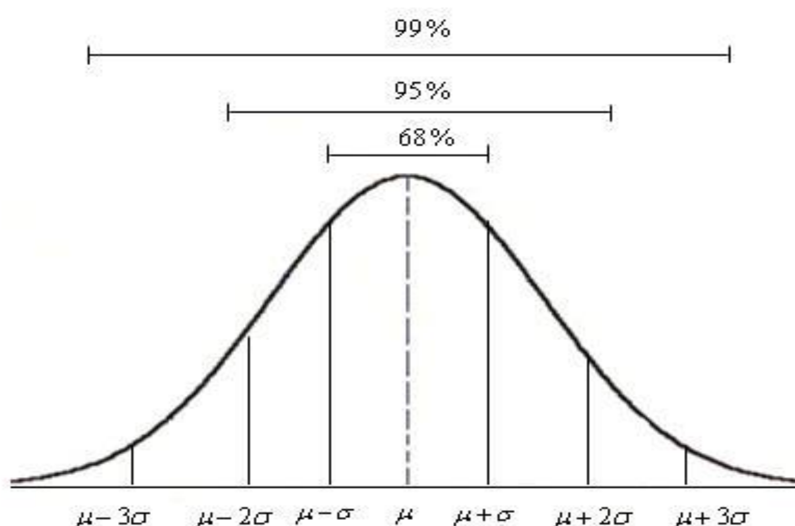


Figura 16 Localización de las observaciones alrededor de la media para una distribución de frecuencia en forma de campana.

Nota:

Para el caso del diagrama de cajas (box plot), entre los límites: Q_1 −1.5*IQR (− 2.698σ) y Q_3 +1.5*IQR (2.698σ), alrededor de la mediana, se encuentra aproximadamente el 99.3% de los datos.

2.6 EL ÍNDICE DE GINI

El **Índice de Gini** es una medida de la desigualdad ideada por el estadístico italiano Corrado Gini. Normalmente se utiliza para medir la desigualdad en los ingresos, pero puede utilizarse para medir cualquier forma de distribución desigual (la tierra, por ejemplo). El Índice de Gini es un número entre 0 y 1, en donde 0 se corresponde con la perfecta igualdad (todos tienen los mismos ingresos) y 1 se corresponde con la perfecta desigualdad (una persona tiene todos los ingresos y todos los demás ninguno). El **índice de Gini** es el coeficiente de Gini expresado en porcentaje, y es igual al coeficiente de Gini multiplicado por 100.

Aunque el coeficiente de Gini se utiliza sobre todo para medir la desigualdad en los ingresos, también puede utilizarse para medir la desigualdad en la riqueza. Este uso requiere que nadie disponga de una riqueza neta negativa. El coeficiente de Gini se calcula como una relación de las áreas en el diagrama de la curva de Lorenz. Si el área entre la línea de perfecta igualdad y la curva de Lorenz es A, y el área por debajo de la curva de Lorenz es B, entonces el coeficiente de Gini es A/(A+B). Esta relación se expresa como porcentaje o como equivalente numérico de ese porcentaje, que es siempre un número entre 0 y 1.
El coeficiente de Gini en varios países (del Informe de Desarrollo Humano de Naciones Unidas 2004):

Namibia:	70.7	Brasil:	59.1
Rusia:	45.6	China:	44.7
EE. UU.	40.8	India:	32.5
Dinamarca:	24.7	Hungría:	24.4

De forma resumida la Curva de Lorenz es una gráfica de concentración acumulada de la distribución de la riqueza superpuesta a la curva de la distribución de frecuencias de los individuos que la poseen y su expresión en porcentajes es el Índice de Gini.

La Curva de Lorenz es un gráfico frecuentemente utilizado para representar la distribución relativa de una variable en un dominio determinado. El dominio puede ser el conjunto de hogares o personas de una región o país, por ejemplo. La variable cuya distribución se estudia puede ser el ingreso de los hogares o las personas. La curva se gráfica considerando en el eje horizontal el porcentaje acumulado de personas u hogares del dominio en cuestión y el eje vertical el porcentaje acumulado del ingreso.

Cada punto de la curva se lee como porcentaje cumulativo de los hogares o las personas. La curva parte del origen (0,0) y termina en el punto (100,100). Si el ingreso estuviera distribuido de manera perfectamente equitativa, la curva coincidiría con la línea de 45 grados que pasa por el origen (por ejemplo el 30% de los hogares o de la población percibe el 30% del ingreso). Si existiera desigualdad perfecta, o sea, si un hogar o persona poseyera todo el ingreso, la curva coincidiría con el eje horizontal hasta el punto (100,0) donde saltaría el punto (100,100).

En general la curva se encuentra en una situación intermedia entre estos dos extremos, si una curva de Lorenz se encuentra siempre por encima de otra (y, por lo tanto, está más cerca de la línea de 45 grados) podemos decir sin ambigüedad que la primera exhibe menor desigualdad que la segunda. Esta comparación gráfica entre distribuciones de distintos dominios geográficos o temporales es el principal empleo de las curvas de Lorenz.

Ejemplo 27

Para ilustrar el cálculo del Índice de Gini y la elaboración de la Curva de Lorenz, se tomarán los datos de volumen de ventas (en millones de euros) de 200 empresas españolas. Notará que en los cálculos se utilizará la metodología de estadística descriptiva para datos agrupados.

Volumen de Ventas Límites										
inferior	superior	fi	Fa	Fr	p = Fra	mi	mifi	MR_i	q = MRA	p−q
50	100	30	30	15.0	15	75	2250	1.13	1.13	13.88
100	200	25	55	12.5	27.5	150	3750	1.88	3.00	24.50
200	500	40	95	20.0	47.5	350	14000	7.00	10.00	37.50
500	1000	50	145	25.0	72.5	750	37500	18.75	28.75	43.75
1000	2000	25	170	12.5	85	1500	37500	18.75	47.50	37.50
2000	5000	30	200	15.0	100	3500	105000	52.50	**100.00**	
Sumatorias		200			247.5		200000	**100.00**		157.13

fi = frecuencia absoluta, representa el número de empresas por clase.
Fa = frecuencia absoluta acumulada
Fr = frecuencia relativa.
Fra = frecuencia relativa acumulada y se representará con la letra **p**.
mi = marca de clase

$$MR_i = \frac{m_i f_i}{\sum_{i=1}^{k} m_i f_i}$$

MRA = son los valores acumulados de MR y se representarán con la letra **q**.

$$IG = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{157.13}{247.5} = 0.63$$

Para construir la Curva de Lorenz, se utiliza la tabla siguiente:

P	q
0	0
15.00	1.13
27.50	3.00
47.50	10.00
72.50	28.75
85.00	47.50
100.00	100.00

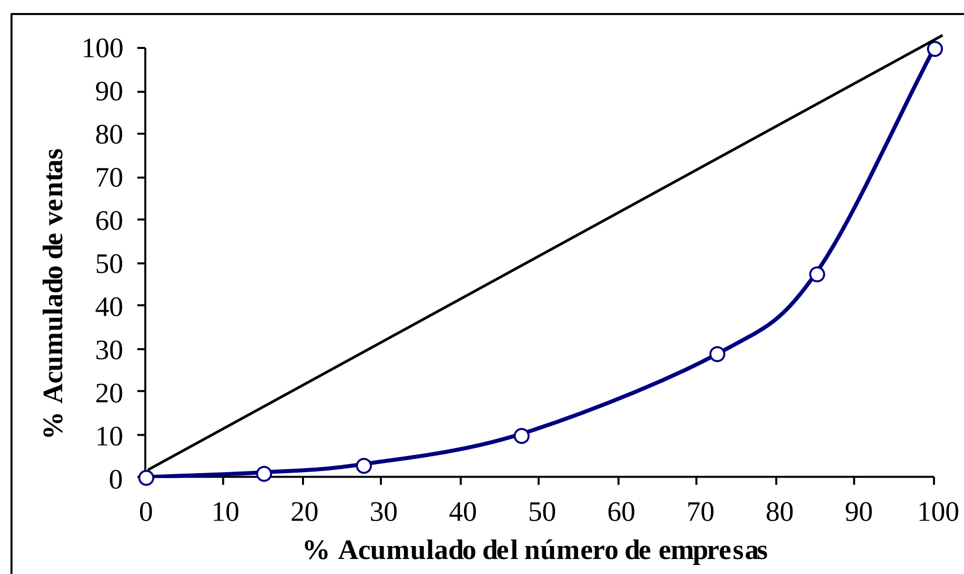


Figura 17 **Curva de Lorenz del volumen de ventas (en millones de euros) de 200 empresas españolas**

La manera de interpretarla será: cuanto más cerca se sitúe esta curva de la diagonal, menor concentración habrá, o más homogeneidad en la distribución. Cuanto más se acerque a los ejes, por la parte inferior del cuadrado, mayor concentración.

2.7 PRESENTACIÓN TABULAR: CUADROS

Los cuadros son presentaciones tabulares que muestran la información de manera ordenada por filas y por columnas, de manera visualmente agradable. Los cuadros estadísticos bien elaborados son muy importantes ya que permiten presentar y divulgar la información obtenida en las operaciones estadísticas, de una manera fácil de interpretar y útil para el usuario.

Desde el punto de vista técnico, Trejos y Moya (1998) citan que un cuadro estadístico se define como: "una lista de datos cuantitativos interrelacionados (es decir, cifras que se aplican a fenómenos concretos y correlacionados en tiempo, lugar, etc., definidos), distribuidos en columnas y filas con palabras, frases y afirmaciones explicativas y aclaratorias, en número suficiente, en forma de títulos, encabezados y notas que aclaren el significado completo de datos y su origen".

Lo esencial en un cuadro estadístico es que la información presentada sea fidedigna, es decir, que la misma sea verdadera y exacta, y que sea legible, es decir, que cualquier lector, sin ser especialista, sea capaz de comprender lo que se está presentando. La principal ventaja de un cuadro es que comunica claramente la información sin necesidad de texto. Un cuadro estadístico está constituido por:

1) Número de cuadro
2) Título
3) Columna matriz
4) Encabezados
5) Cuerpo o contenido
6) Nota introductoria o preliminar
7) Nota al pie
8) Fuente

En la elaboración de cuadros estadísticos no existen reglas formales aceptadas universalmente, aunque sí normas internacionales establecidas por algunos centros especializados en Estadística.

A continuación se brindan algunas sugerencias para la adecuada construcción de cuadros con el fin de estandarizar la presentación de la información estadística en este curso. Éstas se apegan a las sugerencias dadas por la mayoría de autores especializados.

1. Número de cuadro

Se utiliza siempre que haya más de un cuadro dentro del documento donde el mismo se presenta; este número es importante para identificarlo o ubicarlo en una presentación.

2. Título

Es una breve explicación de la naturaleza, clasificación y referencia en el tiempo de los datos presentados. Debe responder a las preguntas: qué son los datos, cuándo y dónde se recolectaron, cómo y bajo qué criterios se clasificaron.

3. Columna matriz

Es la primera columna, a la izquierda del cuadro y ella contiene la clasificación principal que se hizo de la información.

4. Encabezados

Es la parte del cuadro en la que están situados los títulos del resto de las columnas, describiendo en forma general la (s) clasificación (es) de cada columna.

5. Cuerpo o contenido

Es la parte del cuadro que contiene las cifras o datos que se quiere presentar.

6. Nota introductoria o preliminar

Es una frase, generalmente entre paréntesis o guiones, colocada debajo del título. Explica o provee información relacionada con el cuadro, por ejemplo, se puede utilizar para:

- Indicar las unidades en que se trabaja,
- Darle más claridad al título,
- Prevenir al usuario de las limitaciones de la información,
- Establecer la base sobre la que se realizan las comparaciones.

7. Nota al pie

Es una frase que explica o aclara cierta cifra o clasificación, su función es más específica que la de la nota introductoria. Para indicar la nota al pie se utilizan llamadas de atención (números, símbolos como / ó *).

8. Fuente

Es una cita bibliográfica exacta del origen de los datos. No se incluye cunado los datos contenidos en el cuadro fueron obtenidos directamente por la persona o institución que lo confecciona.

A continuación se presenta un ejemplo:

CUADRO 1 **(1)**

(2) TONELADAS MÉTRICA DE AZÚCAR POR HECTÁREA (TAH) Y TONELADAS MÉTRICAS DE CAÑA POR HECTÁREA (TCH) DE VARIEDADES FLOREADORAS DE LA QUINTA PRUEBA REGIONAL DE CENGICAÑA EVALUADAS EN PLANTÍA Y PRIMERA SOCA EN ZONA ALTA. ZAFRAS 2003-04 y 2004-05

(Media general) **(6)**

Variedad **(3)**	TAH **(4)**			TCH		
	Plantía	1ra. Soca	Media de dos cortes	Plantía	1ra. soca	Media de dos cortes
CG98-62	16.13	14.41	15.27	115.1	104.6	109.8
CG98-41	14.95	13.98	14.46	107.6	92.9	100.2
RB72-1012	15.68	13.16	14.42	122.1	95.8	109.0
CG98-46	17.28	11.33	14.31	120.4	84.3	102.4
CP73-1547	14.97	12.42	13.69	106.8	88.5	97.6
PR87-2015	16.13	11.10	13.61	125.4	96.4	110.9
CP96-1967	14.21	12.34	13.27	99.4	85.9	92.6
SP79-1287	14.64	11.62	13.13	109.0	87.7	98.3
CG98-78	13.72	11.96	12.84	104.5	87.4	95.8
CP72-2086	14.71	10.08	12.39	100.4	69.2	84.8

Fuente: CENGICAÑA, Memoria de presentación de resultados de investigación, 2005. **(8)**

2.8 ANÁLISIS BIVARIADO

Puede ocurrir en problemas prácticos, que tengamos interés en estudiar simultáneamente dos o más variables cuantitativas, cualitativas, o ambas.

2.8.1 DESCRIPCIÓN TABULAR

Para hacer un análisis de datos, generalmente se disponen los datos en las llamadas tablas de datos, que son arreglos rectangulares en forma de matriz, en las que las filas y columnas describen a individuos o variables, según sea el caso.

a) Tablas de individuos × variables

Son tablas en las que los individuos están en las filas y las variables en las columnas. Supóngase que se tiene *n* individuos descritos por *p* variables, la tabla de datos quedará de la siguiente forma:

Individuos	Variables					
	X_1	X_2	. . .	X_j	. . .	X_p
1	x_{11}	x_{21}	. . .	x_{j1}	. . .	x_{p1}
2	x_{12}	x_{22}	. . .	x_{j2}	. . .	x_{p2}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.i	x_{1i}	x_{2i}	. . .	x_{ji}	. . .	x_{pi}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
N	x_{1n}	x_{2n}	. . .	x_{jn}	. . .	x_{pn}

Ejemplo 28

Al realizar una encuesta, normalmente se disponen los datos en una tabla de individuos × variables. Considérese que en una encuesta se ha recabado información como el nombre, el sexo, la edad, ele estado civil, el número de hijos, el ingreso mensual bruto, etc. Entonces la tabla de datos tendría una forma como la mostrada a continuación:

Nombre	Sexo	Edad (años)	Estado civil	Número de hijos	Ingreso mensual (Q)
J. Pérez	M	34	Casado	1	5600
S. Velásquez	F	24	Soltera	0	3800
L. Mérida	F	52	Viuda	3	5100
F. Solís	M	46	Soltero	0	6000
A. Flores	F	38	Casada	2	6200
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

b) Tablas de variable × variables

Se trata de tablas en que tanto las filas como las columnas describen a variables, o a modalidades de éstas, en el caso cualitativo. Las tablas usadas en este caso son conocidas como: *tablas de doble entrada, tablas de asociación, tablas de contingencia o distribuciones conjuntas de frecuencias.* Estas tablas presentan dos márgenes, cada cual con los totales referentes a una de las variables.

Ejemplo 29

En la tabla 1 se presentan las variables X = conceptos en el curso de Estadística, con modalidades, x_1 = Deficiente (D), x_2 = Regular (R), x_3 = Bueno (B) y x_4 = Excelente (E), y Y = carrera universitaria frecuentada, con las modalidades, y_1 = Agronomía (A) y y_2 = Veterinaria (V).

Tabla 3 Distribución de los alumnos de la Universidad UFSC, según el concepto en Estadística y la carrera universitaria. (*)

Carrera (Y)	Concepto obtenido (X)				Total por carrera
	Deficiente (D)	Regular (R)	Bueno (B)	Excelente (E)	
Agronomía	10	60	50	0	120
Veterinaria	40	40	20	20	120
Total por concepto	50	100	70	20	240

(*) Otro título para esta tabla podría ser: "Distribución conjunta de frecuencias de las variables concepto en Estadística y carrera universitaria"

Obsérvese que la línea de totales ofrece la distribución de la variable Y, así como la columna de los totales ofrece la distribución de la variable X. Las distribuciones así obtenidas son llamadas técnicamente de: *Distribuciones marginales*.

Note también que con las frecuencias marginales de la distribución conjunta, se pueden rescatar las distribuciones simples de las variables X y Y. En ese contexto, una distribución simple siempre puede ser vista como distribución marginal de alguna distribución conjunta.

Tabla 4a Distribución marginal de X (Conceptos obtenidos)

Concepto (x_i)	Frecuencia (f_i)
Deficiente	50
Regular	100
Bueno	70
Excelente	20
Total	240

Tabla 4b Distribución marginal de Y (Carrera universitaria)

Carrera universitaria (y_j)	Frecuencia (f_j)
Agronomía	120
Veterinaria	120
Total	240

Con base en la distribución conjunta de las frecuencias absolutas, se puede construir las distribuciones condicionales de frecuencias absolutas de X para un dado valor de Y o, de Y para un dado valor de X. La tabla 5a se refiere a la distribución condicional de frecuencias absolutas de X | y =A o equivalentemente, de X | y_1, que se interpreta como "distribución condicional de frecuencias absolutas de los conceptos obtenidos en Estadística *dado que* el curso es de Agronomía". De modo análogo, la tabla 5b muestra la distribución condicional X | y =V o equivalentemente, de X | y_2. Por otra parte, las tablas 6a hasta 6d ofrecen las distribuciones condicionales de frecuencias absolutas de las carreras universitarias *dado* cada concepto, para lo cual se tiene respectivamente Y | x_1, Y | x_2, Y | x_3 y Y | x_4.

Tabla 5a Distribución condicional de X | y_1 (Distribución de los alumnos de Agronomía según sus conceptos en Estadística.

Concepto (x_i)	Frecuencia (f_i)
Deficiente	10
Regular	60
Bueno	50
Excelente	0
Total	120

Tabla 5b Distribución condicional de X | y_2, (Distribución de los alumnos de Veterinaria según sus conceptos en Estadística.

Concepto (x_i)	Frecuencia (f_i)
Deficiente	40
Regular	40
Bueno	20
Excelente	20
Total	120

Tabla 6a Distribución condicional de Y | x_1 (Distribución de los alumnos deficientes en Estadística, según la carrera universitaria.

Carrera (y_j)	Frecuencia (f_j)
Agronomía	10
Veterinaria	40
Total	50

Tabla 6b Distribución condicional de Y | x_2, (Distribución de los alumnos regulares en Estadística, según la carrera universitaria.

Concepto (y_i)	Frecuencia (f_j)
Agronomía	60
Veterinaria	40
Total	100

Tabla 6c Distribución condicional de Y | x_3 (Distribución de los alumnos buenos en Estadística, según la carrera universitaria.

Carrera (y_j)	Frecuencia (f_j)
Agronomía	50
Veterinaria	20
Total	70

Tabla 6d Distribución condicional de Y | x_4, (Distribución de los alumnos excelentes en Estadística, según la carrera universitaria.

Concepto (y_i)	Frecuencia (f_j)
Agronomía	0
Veterinaria	20
Total	20

Generalizando, la Tabla 3, de doble entrada, se representa de la siguiente forma:

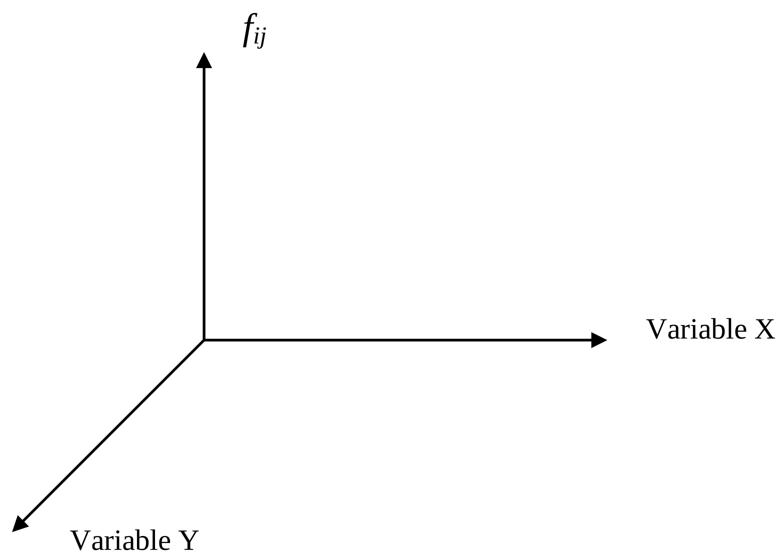
Modalidades de y	Modalidades de x						Total
	1	2	. . .	K	. . .	q	
1	f_{11}	f_{21}	. . .	f_{k1}	. . .	f_{q1}	$f_{.1}$
2	f_{12}	f_{22}	. . .	f_{k2}	. . .	f_{q2}	$f_{.2}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
J	f_{1j}	f_{2j}	. . .	f_{kj}	. . .	f_{qj}	$f_{.j}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
P	f_{1p}	f_{2p}	. . .	f_{kp}	. . .	f_{qp}	$f_{.p}$
Total	$f_{1.}$	$f_{2.}$	. . .	$f_{k.}$	. . .	$f_{q.}$	$f_{..} = n$

2.8.2 DESCRIPCIÓN GRÁFICA

a) Variables cualitativas bidimensionales

De modo general, los gráficos de las variables cualitativas bidimensionales son de dos tipos:

- Gráficos tridimensionales
  Son compuestos de paralelogramos, separados entre sí, descritos en ejes tridimensionales:



Sus bases son determinadas por los valores de las variables X y Y, y su altura por las frecuencias f_{ij}. Para los datos del ejemplo 13 tenemos el siguiente gráfico:

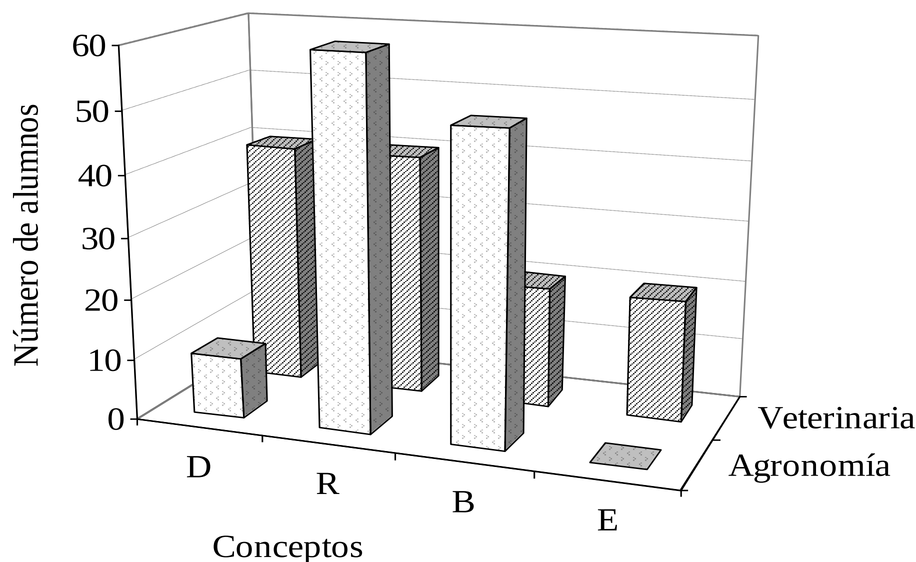


Figura 18 **Distribución de los alumnos según el concepto en Estadística y la carrera universitaria cursada.**

- Gráficos de distribuciones condicionales

Estos tipos de gráficos pueden simplificar la descripción de las variables cualitativas bidimensionales. Dos gráficos, en general, pueden ser construidos, uno con las distribuciones condicionales X | y_j , que proveerá la distribución de los conceptos "dentro" de cada curso y, otro con las distribuciones condicionales de Y | x_i , que exhibe la distribución de los cursos "dentro" de cada concepto. Naturalmente, uno de ellos será escogido, de acuerdo con el interés del usuario.

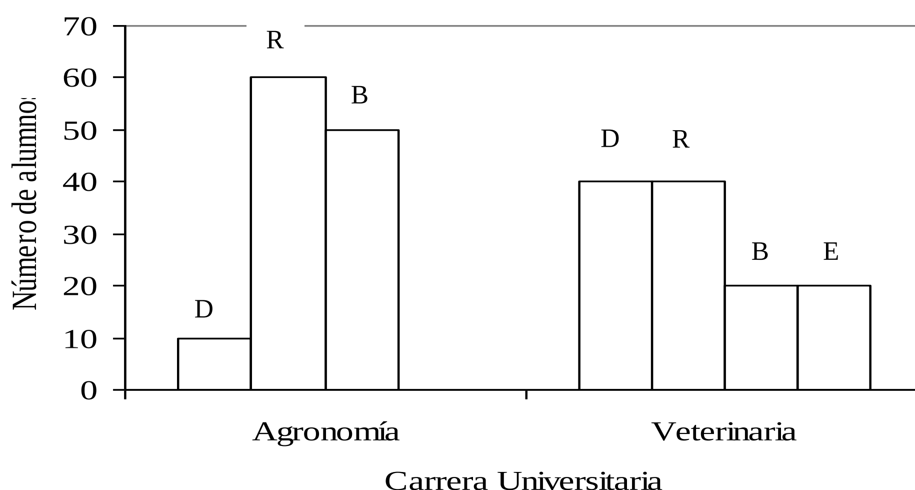


Figura 19 **Distribución de los alumnos, en cada curso, según sus conceptos en Estadísticas**

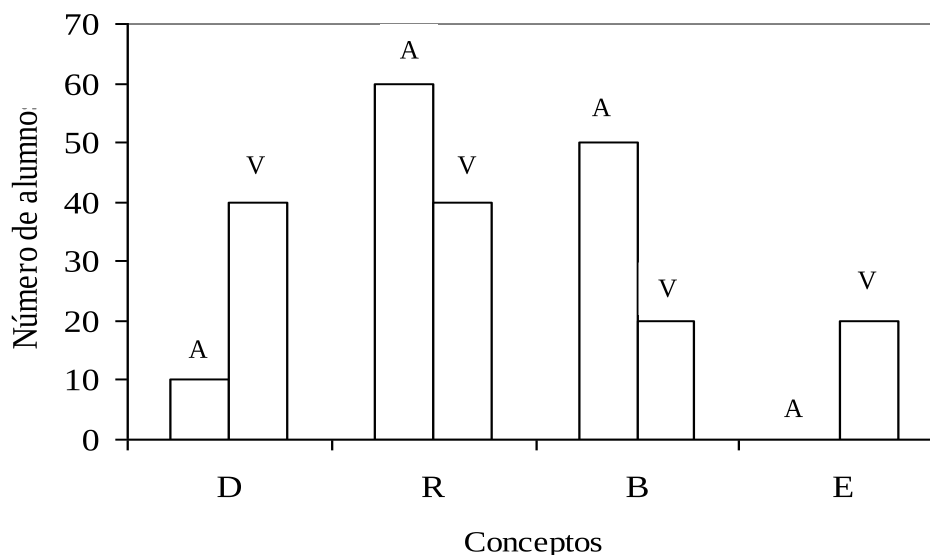


Figura 20 **Distribución de los alumnos, en cada concepto, según la carrera universitaria cursada.**

b) Variables cuantitativas bidimensionales

Si las dos variables X y Y son cuantitativas, se suele estudiar su relación dibujando un *diagrama de dispersión*. Este diagrama permite visualizar los valores de las dos variables, pues grafica la forma de la forma de la nube de puntos constituidos por las parejas de datos $\{(x_i, y_i)/i=1, 2, \ldots, n\}$. La forma del diagrama de dispersión dará una idea de la relación que pueda existir entre las dos variables.

Ejemplo 30

Los datos de la siguiente tabla de correlación, se refieren a los pesos (en kg) al nacer (X) y peso (en kg) al destete (Y) de n=12 becerros de la raza Guzerat.

X	25.3	26.9	26.5	27.4	27.9	25.8	28.4	28.9	27.6	27.2	27.5	28.1
Y	48.4	49.7	49.2	50	50.6	48.7	51.6	52.3	50.4	50	50.7	50.9

El diagrama de dispersión tiene el siguiente aspecto:

52.49
51.42
50.35
49.28
48.20
Peso al destete (kg)
25.12 26.11 27.10 28.09 29.08
Peso al nacer (kg)

Figura 21 Diagrama de dispersión

2.8.3 COVARIANZA

La covarianza proporciona una idea del signo y de la cantidad de relación entre dos variables, a través de su variación conjunta. Así, por ejemplo, la covarianza es positiva si la relación es de tipo directo, esto es, si a pequeños valores de X corresponden pequeños valores de Y. De modo análogo, ella es negativa, si la relación es de tipo inverso, o sea, para pequeños valores de una de las variables, corresponden grandes valores de la otra, y viceversa. Además, ella será nula o próxima de cero para relaciones lineares débiles.

Esto en general ocurre cuando los puntos están dispersos en torno de las rectas $x = \bar{x}$ y $y = \bar{y}$.

Definición: Se define covarianza entre las variables cuantitativas X y Y por la siguiente expresión:

$$\hat{Cov}[X, Y] = \frac{1}{N-1}\sum_{i=1}^{N}\left[(x_i - \mu_X)(y_i - \mu_Y) \right]$$

Su estimador muestral es:

$$\hat{Cov}[X, Y] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i\, y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \right]$$

Ejemplo 31

Los datos de la tabla que se presentan a continuación, se refieren a los pesos de los padres (X) y de sus hijos (Y) en kilogramos.

	x_i	y_i
1	78	60
2	65	52
3	86	68
4	68	53
5	83	65
6	68	57
7	75	58
8	80	62
9	82	65
10	66	53

$$\hat{\text{Cov}}[X,Y] = \frac{1}{10-1}\left[44921 - \frac{(751)(593)}{10}\right] = 42.97$$

La covarianza es de utilidad para indicar el signo de la relación entre dos variables. Así, en el ejemplo anterior $\hat{\text{Cov}}[X,Y] = 42.97$, muestra que la relación entre X y Y es de tipo positivo o directo, esto es, para grandes valores de X, corresponden grandes valores de Y y viceversa. Sin embargo, en lo referente a la cuantificación de la relación, ¿el valor 42.97 refleja un pequeño o alto grado de relación? No existe un valor que pueda ser usado como referencia para saber lo que es una grande o una pequeña covarianza.

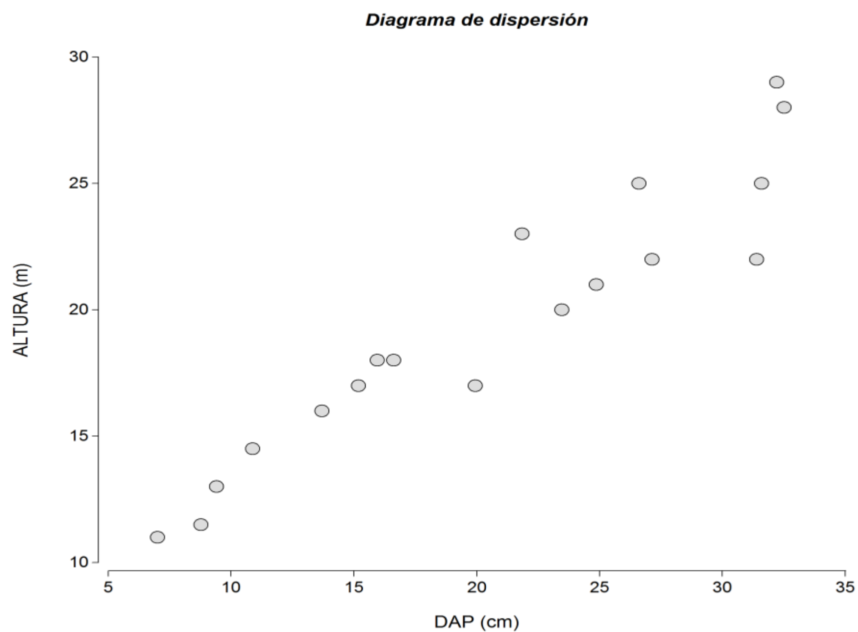
Para tratar este problema, Karl Pearson propuso una nueva medida para medir la relación entre variables cuantitativas: el coeficiente de correlación (tema a tratar en la Unidad VI)

Ejemplo 32

Calcule el valor de la Covarianza y construya un diagrama de dispersión entre las variables DAP (cm) y Altura (m) de 18 árboles de *P. maximinoii* H.E. Moore

DAP	ALTURA
7.01	11.00
8.77	11.50
9.41	13.00
10.87	14.50
13.69	16.00
15.18	17.00
15.95	18.00
16.62	18.00
19.94	17.00

DAP	ALTURA
21.85	23.00
23.46	20.00
24.86	21.00
26.60	25.00
27.14	22.00
31.40	22.00
31.60	25.00
32.21	29.00
32.51	28.00



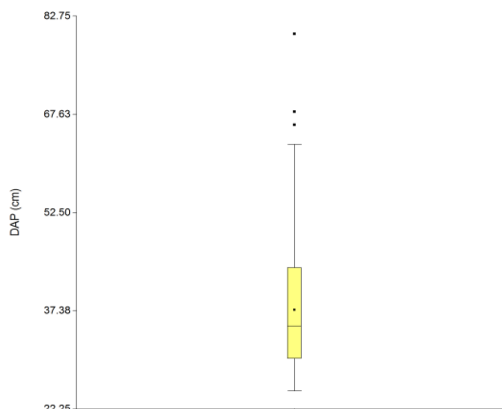
I	x_i	y_i	x_iy_i
1	7.01	11.0	77.11
2	8.77	11.5	100.86
3	9.41	13.0	122.33
4	10.87	14.5	157.62
5	13.69	16.0	219.04
6	15.18	17.0	258.06
7	15.95	18.0	287.10
8	16.62	18.0	299.16
9	19.94	17.0	338.98
10	21.85	23.0	502.55
11	23.46	20.0	469.20
12	24.86	21.0	522.06
13	26.6	25.0	665.00
14	27.14	22.0	597.08
15	31.4	22.0	690.80
16	31.6	25.0	790.00
17	32.21	29.0	934.09
18	32.51	28.0	910.28
Sumatoria	369.07	351.0	7941.31

$$\hat{Cov}[X,Y]=\frac{1}{18-1}\left[7491.31-\frac{(369.07)(351)}{18}\right]=43.79$$

LISTA DE EJERCICIOS 2

1. Al analizar los datos del ejercicio 25, considerándolos como datos sin agrupar, Infostat v.2017 proporcionó la siguiente salida:

Resumen	DAP
n	280.00
Media	37.50
D.E.	9.72
Var(n-1)	94.45
CV	25.92
Mín	25.00
Máx	80.00
Mediana	35.00
Q1	30.00
Q3	44.00
Asimetría	1.22
Kurtosis	2.11



Se le solicita:

a) Interprete las medidas de tendencia central y de dispersión para datos sin agrupar. Compare y discuta los resultados para datos agrupados.

b) Usando MS Excel, calcule el valor de la o las modas.

c) En el *box plot* señale sus partes, indique el tipo de asimetría, señale los puntos atípicos.

d) Indique el tipo de simetría y curtosis que posee la distribución del DAP.

2. Los datos de producción de resina (expresados en kilogramos) de 40 árboles de *Pinus elliotti* se presentan en la siguiente tabla:

0.71	1.53	1.94	2.16	2.39	2.67	3.06	3.34	3.57	3.93
0.75	1.57	2.04	2.16	2.48	2.75	3.09	3.37	3.63	3.94
1.20	1.67	2.06	2.18	2.48	2.77	3.26	3.55	3.69	4.05
1.42	1.80	2.06	2.22	2.63	2.78	3.32	3.56	3.77	5.41

Con estos datos (considerándolos sin agrupar):

a) Calcule las medidas de tendencia central y de dispersión. Discuta los resultados.

b) Construya un *box plot*.

c) Determine el tipo de simetría y de curtosis que posee el conjunto de datos.

d) Construya un gráfico de tallos y hojas.

3. En la tabla siguiente se presentan los datos de diámetro a la altura del pecho (cms) de 30 árboles de tres especies de pino: *Pinus strobus* L. var. chiapensis (A), *P. maximinoii* H.E. Moore (B) y *P. caribaea* Morelet (C), muestreados en el proyecto de reforestación Saquichaj, Cobán (Alta Verapaz). Los datos fueron tomados de la tesis de Gerardo Paíz Schwartz (1998):

P. strobus	*P. maximinoii*	*P. caribaea*
8.70	7.01	11.79
9.21	8.12	12.34
11.21	8.70	14.95
11.71	8.93	15.00
14.22	9.41	15.05
14.67	9.50	15.40
15.24	10.87	15.55
15.36	12.08	16.94
15.80	13.70	17.71
16.13	14.35	17.83
17.70	15.20	18.69
17.92	15.30	19.51
18.33	15.95	20.80
18.98	16.10	20.90
21.57	16.62	21.01
23.70	18.68	21.84
24.40	19.94	21.85
25.60	21.56	22.67
26.23	21.85	23.05
26.25	23.00	25.82
27.03	23.46	26.00
28.00	23.64	26.23
28.36	24.86	26.45
30.02	25.30	26.60
30.46	26.60	27.46
31.48	26.68	27.50
31.60	27.14	27.83
31.92	28.75	28.10
32.21	30.20	30.85
32.51	31.45	30.88

Con estos datos:

a) Calcule las medidas de tendencia central y de dispersión para datos sin agrupar. Compare y discuta los resultados para las tres especies.

b) Construya un *box plot* para cada especie, preséntelos en una misma figura, compare los resultados.

c) Determine el tipo de simetría que posee cada una de las distribuciones de DAP.

4. Un profesor decide utilizar un promedio ponderado para obtener las calificaciones finales de los estudiantes que acuden al curso que imparte. El promedio de tareas tendrá un valor de 20% de la calificación del estudiante; el examen semestral, 25%; el examen final, 35%; el artículo de fin de semestre, 10%, y los exámenes parciales, 10%. A partir de los datos siguientes calcule el promedio final para los cinco estudiantes del curso.

Estudiante	Tareas	Parciales	Artículo	Ex. semestral	Ex. Final
1	85	89	94	87	90
2	78	84	88	91	92
3	94	88	93	86	89
4	82	79	88	84	93
5	95	90	92	82	88

5. Un empresario se encuentra calculando el factor de crecimiento promedio de su almacén de aparatos de sonido en los últimos seis años. Utilizando una media geométrica, llega a un resultado de 1.24. Los factores de crecimiento individuales de los últimos cinco años fueron 1.19, 1.35, 1.23, 1.19 y 1.30. Pero el empresario perdió los registros del sexto año después de haber calculado la media. ¿Cuál era el factor de crecimiento del último año?

6. El peso medio de 5 cajas de tomate fue 433 kg con desviación estándar igual a 18.235 kg. Si son adicionadas 3 cajas con pesos de 400, 480 y 375 kg calcule la nueva media y nueva desviación estándar.

7. Un conjunto de 60 observaciones posee una $\bar{x}$ = 66.8 y s^2 = 12.60 y una forma de distribución desconocida.

a) ¿Entre qué valores deberán estar al menos 75% de las observaciones, de acuerdo con el teorema de Tchebychev?

b) Si la distribución es simétrica, aproximadamente ¿cuántas observaciones deberán encontrarse en el intervalo 59.7 a 73.9?

8. De una población de proveedores de caña de azúcar del ingenio azucarero "Santa Teresinha" en Piracicaba, Estado de São Paulo, referente a la zafra 1990, se retiró aleatoriamente una muestra de tamaño igual a 30 proveedores, cuya producción en tm/ha (y_j) es dada a continuación:

88	121	81	125	119	101
87	77	183	76	85	85
92	100	127	59	79	74
113	97	101	90	88	74
129	79	81	128	108	96

a) Obtenga los errores (o desvíos) del j-ésimo valor observado con relación a la media ($\bar{y}$), que son dados por la siguiente expresión:

$$\hat{e}_j = y_j - \bar{y} \quad con \quad j = 1, 2, \ldots, n$$

siendo y_j = valor observado referente al j-ésimo elemento perteneciente a la muestra en estudio.

b) Verifique que $\sum_{j=1}^{30} \hat{e}_j = 0$

c) Obtenga las estimaciones de la varianza (s^2) y desviación estándar (s) para los datos, sin agrupar.

d) Obtenga la estimación de la varianza $\hat{V}(\bar{y})$ y el error estándar $s_e(\bar{y})$ de la media.

$$\hat{V}(\bar{y}) = \frac{s^2}{n} \qquad\qquad s_e(\bar{y}) = \frac{s}{\sqrt{n}}$$

e) Construya un diagrama de cajas de dispersión.

9. Elabore un diagrama de tallos y hojas para los resultados del número de adultos de chinche salivosa que se localizaron en las macollas de caña de azúcar, muestreadas en la finca "Nuevo Mundo".

57	17	24	40	21	17	37	15	13
25	19	45	19	21	42	26	16	48
37	48	17	26	50	14	27	56	25
55	15	21	59	13	19	43	12	28

10. Los rendimientos comerciales de frutos de tomate (expresados en kg/ha) en 24 parcelas experimentales localizadas en el Valle de Salamá, son presentados en la siguiente tabla:

29.1	33.4	28.5	39.5	38.1	30.8
38.5	24.4	41.4	29.8	25.9	38.9
33.4	30.4	35.3	24.9	33.8	29.4
29.7	39.1	30.5	30.5	35.5	31.6

Calcule:

a) El valor de la producción que separa el 25% de las parcelas con tomate más productivas.

b) Calcular el percentil de orden 7.5

11. A continuación se presentan tres diagramas de tallos y hojas, a partir de los cuales, se le solicita reconstruir los datos originales.

11.1

```
Stem Leaf                 # (frecuencia)
 6 2                      1
 5
 5 4                      1
 4
 4 1                      1
 3 566677899              9
 3 1334                   4
 2 5567889                7
 2 01122334               8
 1 56789                  5
 1 24                     2
 0 78                     2
   ----+----+----+----+
Multiply Stem.Leaf by 10**0
```

11.2
```
Stem Leaf                #
  6 2                    1
  5
  5 4                    1
  4
  4 1                    1
  3 566677899            9
  3 1334                 4
  2 5567889              7
  2 01122334             8
  1 56789                5
  1 24                   2
  0 78                   2
    ----+----+----+----+
Multiply Stem.Leaf by 10**-1
```

11.3
```
Stem Leaf                #

  6 1                    1
  5 6                    1
  5 3                    1
  4 78                   2
  4 01                   2
  3
  3 03                   2
  2 5                    1
  2 002244               6
  1 566889               6
  1 02233334             8
   ----+----+----+----+
Multiply Stem.Leaf by 10**+2
```

12. El diámetro de algunos árboles de dos tipos de bosque fue medido en un inventario forestal, obteniéndose los siguientes valores:

BOSQUE A

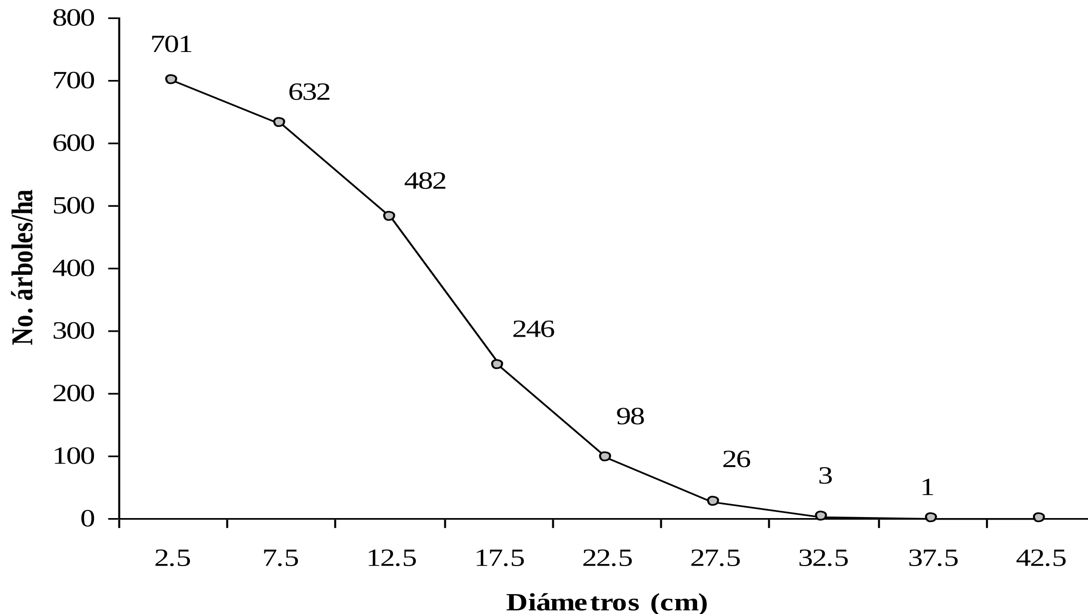
16	50	13	8	5	77	93	27	57	28	24	16
49	60	7	5	9	30	8	51	41	33	62	35
9	49	31	107	27	56	26	55	10	18	7	24
17	63	11	34	19	12	40	28	6	19	10	50
16	29	22	10	17	36	42	134	7	10	29	14
12	12	29	76	10	106	52	43	17	16	51	19
21	96	87	29	77	6	9	21	18	6	15	161
32	12	16	29	7	20	37	76	47	6	17	35
30	44	13	56	112	38	15	56	17	34	43	6
52	42	35	25	31	127	9	21	5	154	13	7

BOSQUE B

38	43	32	18	47	33	38	27	50	34	34	31
28	31	46	27	33	33	38	24	33	23	16	42
22	26	27	32	23	46	30	9	36	47	21	61
34	37	36	30	41	16	7	33	50	11	27	7
23	27	38	23	25	33	30	36	27	32	23	25
30	23	40	15	23	47	35	39	41	46	35	30
42	23	43	35	28	31	35	33	30	30	49	34
48	29	29	30	21	32	28	31	36	22	26	41
33	25	42	30	22	17	40	39	36	55	29	40
42	29	37	29	32	49	17	22	48	31	42	38

a) Calcule las medidas de tendencia y de dispersión para cada conjunto de datos.
b) Construya un histograma y un polígono de frecuencias para cada conjunto de datos.
c) ¿Existen observaciones extremas en esos conjuntos de datos? (construya un *box plot* para los datos de cada bosque)
d) Describa la forma de la distribución en los dos tipos de bosque.

13. En una comunidad de Cobán, Alta Verapaz, se realizó un inventario forestal a un bosque de *Cupressus lusitanica* Miller. Una de las variables medidas fue el diámetro de los árboles. A continuación se presenta una Ojiva de Galton de tipo "mayor que", que se construyó con estos datos:



Con esta información:

a) Construya la tabla de frecuencias y calcule las medidas de tendencia central y de dispersión.
b) Construya un histograma y analice la simetría de la distribución de datos.

14. A continuación se presenta la distribución por clase diamétrica del número de árboles por hectárea de dos especies de pino: *P. maximinoii* H.E. Moore y *P. caribaea* Morelet:

Clase diamétrica	*P. maximinoii*	*P. caribaea*
[5 – 10)	55	33
[10 – 15)	132	149
[15 – 20)	163	206
[20 – 25)	139	75
[25 – 30)	70	19
[30 – 35)	33	5
[35 – 40)	8	1

Calcule las medidas de tendencia central y de dispersión, compare y analice ambas distribuciones. Además construya un histograma y un polígono de frecuencias para ambas distribuciones.

15. A continuación se presentan los valores de precipitación pluvial mensual (mm), registrados en la Estación convencional de la Escuela Superior de Agricultura "Luiz de Queiroz" localizada en la ciudad de Piracicaba, Estado de Sao Paulo (Brasil), Latitud de 22 42' 30" sur y Longitud de 47 38' 00" oeste, a una altitud de 546 metros sobre el nivel del mar.

Año	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Total
1917	295.7	135.8	58.9	116.2	58.5	13.0	13.3	5.4	62.2	58.4	19.8	194.3	1031
1918	291.0	81.3	152.1	25.2	22.9	15.8	55.1	26.9	68.6	130.9	125.9	127.2	1123
1919	262.8	96.2	30.6	78.0	38.9	170.3	0.0	54.6	44.6	68.2	142.0	102.4	1089
1920	209.9	135.3	164.0	50.2	12.3	16.0	42.9	16.8	111.3	82.3	217.6	161.6	1220
1921	265.7	101.5	17.8	31.7	26.5	25.3	3.8	26.5	35.2	59.7	37.7	180.6	812
1922	387.2	176.6	117.2	29.5	17.4	72.6	22.0	76.2	7.7	91.7	109.7	105.6	1213
1923	219.7	144.4	311.4	89.8	82.9	114.7	5.7	31.5	200.5	41.7	60.7	125.5	1429
1924	108.5	154.2	139.5	15.4	53.9	19.0	0.0	13.0	15.5	15.2	128.4	231.8	894
1925	117.9	55.2	84.7	40.4	59.0	73.5	33.0	0.0	42.7	110.7	230.3	159.4	1007
1926	237.0	20.6	214.4	238.5	33.7	32.0	139.2	59.8	69.0	67.6	74.4	360.8	1547
1927	200.6	259.4	189.6	22.5	0.0	31.0	6.0	63.1	257.4	126.1	48.8	109.5	1314
1928	195.5	220.6	155.9	34.2	61.6	66.3	29.3	1.7	0.0	58.8	83.2	272.0	1179
1929	490.9	331.6	32.6	68.2	119.8	4.8	11.8	50.2	33.0	13.4	99.2	303.0	1559
1930	257.0	427.8	23.4	41.0	14.0	22.0	68.5	47.1	52.6	116.4	108.0	328.6	1506
1931	324.2	334.4	184.3	103.0	37.0	16.0	38.9	11.0	138.2	65.8	214.0	149.1	1616
1932	204.2	95.7	124.8	30.4	98.2	44.0	5.6	46.2	48.6	113.4	174.8	336.4	1322
1933	119.6	150.0	41.6	4.0	77.0	53.8	7.0	15.4	50.8	109.2	40.1	302.0	971
1934	220.0	103.4	46.7	21.6	0.0	37.5	0.0	6.0	59.8	70.4	122.9	415.9	1104
1935	91.2	194.4	140.0	68.8	6.8	53.8	16.4	18.4	223.8	211.4	82.0	184.7	1292
1936	84.8	203.4	202.4	37.8	26.6	2.0	20.8	106.4	88.0	51.8	86.8	293.6	1204
1937	139.0	95.8	206.2	203.0	74.4	46.0	0.0	67.0	13.8	155.6	121.8	141.8	1264
1938	158.3	120.6	107.0	66.0	85.5	0.0	13.0	32.5	80.3	171.0	129.6	209.6	1173
1939	275.0	279.9	169.9	49.7	129.8	58.7	20.4	0.0	19.3	41.9	198.9	259.0	1503
1940	220.8	461.7	96.8	18.0	15.1	4.2	1.1	3.2	28.1	95.8	196.1	198.0	1339
1941	215.1	128.7	84.5	56.2	15.2	49.2	15.3	17.6	190.4	115.7	217.0	186.6	1292
1942	136.6	197.2	206.7	84.3	21.8	46.4	57.5	4.0	22.8	24.3	171.3	201.1	1174
1943	325.6	141.4	269.8	37.1	1.2	33.0	0.7	13.5	72.9	215.8	163.3	359.9	1634
1944	118.7	215.2	182.9	23.0	0.9	13.1	2.7	0.0	13.3	183.5	260.1	158.7	1172
1945	322.0	256.7	85.5	120.5	14.3	187.1	29.6	1.6	29.2	126.4	232.9	162.7	1569
1946	202.9	192.6	112.8	18.1	16.9	19.6	88.9	0.3	21.3	96.2	133.4	82.0	985
1947	352.7	338.9	130.7	10.2	39.1	15.0	48.1	68.6	124.0	90.1	114.6	219.6	1552
1948	313.5	247.7	104.3	38.6	52.6	4.1	40.0	15.4	9.1	147.5	147.0	109.0	1229
1949	210.8	206.5	183.3	54.6	49.0	45.0	0.0	5.8	15.8	100.5	45.1	420.5	1337
1950	231.0	321.7	161.1	121.6	8.6	47.9	13.0	0.0	11.1	190.2	179.8	134.0	1420
1951	284.7	215.4	160.4	22.2	8.3	15.5	8.0	33.2	2.3	128.6	222.6	91.0	1192
1952	207.0	218.6	102.4	52.5	0.0	96.9	4.8	7.0	40.7	105.4	117.9	153.6	1107
1953	251.6	107.4	161.8	74.4	27.7	24.3	14.9	23.2	75.6	97.7	82.3	152.4	1093
1954	189.4	204.2	136.6	30.5	99.1	69.7	8.7	0.0	23.2	87.3	39.7	193.0	1081
1955	272.8	63.7	194.5	120.0	57.3	26.5	10.9	86.5	1.0	64.8	127.6	197.3	1223
1956	60.8	136.9	47.1	90.6	114.5	70.2	53.3	74.0	72.7	72.6	23.1	137.0	953
1957	313.6	128.3	185.6	66.9	4.2	20.1	90.6	51.7	123.4	64.2	57.1	197.3	1303
1958	295.6	143.8	167.6	128.2	172.4	70.1	15.2	3.1	96.4	121.2	124.5	151.2	1489
1959	258.4	162.5	196.2	45.2	29.4	19.3	0.1	72.3	23.2	79.6	246.7	191.1	1324
1960	371.9	332.8	74.6	38.4	80.6	63.3	0.0	22.3	8.7	123.1	93.7	314.2	1524
1961	135.5	195.5	129.6	101.7	34.5	14.6	0.1	11.9	2.2	43.4	138.3	157.9	965
1962	132.3	246.0	329.0	17.2	30.8	51.7	20.5	42.0	32.2	198.9	97.3	201.6	1400
1963	308.4	110.6	56.0	5.0	4.2	0.0	0.0	7.6	3.6	148.7	163.4	148.4	956
1964	69.4	202.6	11.8	10.7	44.0	15.0	77.0	18.2	61.3	161.8	52.7	268.4	993

Año	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Total
1965	326.6	278.4	147.5	55.3	78.3	27.8	50.9	3.8	51.3	112.1	99.1	292.0	1523
1966	252.1	198.8	55.6	18.2	43.7	0.0	12.1	27.8	128.7	130.4	85.2	225.4	1178
1967	272.2	196.6	152.1	3.6	4.2	81.5	13.4	0.0	88.7	210.5	77.5	154.9	1255
1968	322.5	91.5	45.5	32.7	14.0	27.1	17.6	36.9	16.3	121.4	62.8	235.5	1024
1969	126.1	40.9	173.8	53.1	36.6	30.0	15.5	31.1	59.3	83.4	212.7	198.9	1061
1970	263.8	293.4	96.5	83.2	75.1	40.2	10.5	98.7	90.6	83.7	83.9	138.8	1358
1971	180.6	81.3	173.5	28.3	91.4	109.3	35.6	3.3	35.9	151.4	34.2	178.7	1104
1972	255.1	238.7	92.1	101.5	64.0	5.3	145.1	51.6	38.8	181.6	122.8	91.8	1388
1973	121.1	115.2	83.5	88.0	36.0	35.7	59.6	19.4	55.5	174.6	121.7	328.6	1239
1974	147.1	60.1	301.3	21.4	6.6	115.8	0.0	5.0	24.0	126.6	168.4	247.7	1224
1975	168.1	357.3	32.9	37.3	11.3	0.0	39.6	32.0	55.1	98.0	236.8	315.1	1383
1976	295.2	287.0	140.9	19.2	46.4	61.8	77.7	61.4	124.8	168.3	127.4	191.4	1602
1977	322.7	37.3	199.8	107.7	0.9	46.8	6.7	19.7	91.4	66.7	134.8	238.1	1273
1978	112.6	111.6	85.1	15.5	74.0	40.6	78.9	4.6	57.2	58.2	138.9	96.6	874
1979	122.4	114.4	84.7	63.9	96.9	0.0	28.0	81.4	98.6	101.4	102.8	105.7	1000
1980	147.6	163.3	42.3	165.3	13.4	36.3	0.0	22.6	62.5	52.5	145.2	185.1	1036
1981	290.1	71.2	65.5	74.5	44.9	68.5	1.6	3.6	5.9	231.2	167.2	211.2	1235
1982	260.0	221.8	132.3	125.7	62.7	162.7	24.5	44.7	8.1	254.5	210.2	215.0	1722
1983	207.5	283.9	225.6	189.5	334.7	174.0	24.0	2.2	197.4	129.1	88.4	161.4	2018
1984	91.5	18.7	47.1	51.5	45.9	0.0	8.9	112.1	90.4	27.0	201.1	196.0	890
1985	131.8	118.6	203.1	134.5	42.6	22.5	1.5	21.8	76.4	11.3	134.3	71.7	970
1986	143.0	96.2	320.8	47.5	88.6	0.0	10.5	133.2	36.7	52.9	110.4	282.3	1322
1987	342.9	203.1	105.6	70.6	199.9	122.7	10.7	5.9	72.9	67.0	113.3	190.5	1505
1988	185.3	328.3	194.3	140.0	104.3	26.8	0.0	0.0	4.1	196.9	86.0	200.4	1466
1989	370.8	298.3	99.0	103.9	33.0	47.0	85.2	32.0	36.1	46.3	99.9	195.1	1447
1990	271.4	145.9	240.2	37.2	47.1	12.7	134.6	40.9	61.0	125.4	124.4	61.1	1302
1991	266.6	217.6	431.9	129.6	42.1	35.1	15.2	4.7	68.9	70.5	56.8	194.5	1534
1992	99.2	72.4	232.7	79.9	73.8	0.6	36.6	11.1	87.5	242.1	232.3	142.2	1310
1993	179.9	232.1	155.9	57.7	119.7	47.0	13.7	51.8	154.4	73.4	91.6	146.4	1324
1994	134.8	154.4	222.7	114.9	56.2	36.7	24.7	0.0	0.5	126.8	249.2	260.8	1382
1995	197.3	415.6	194.6	85.7	64.2	23.5	59.1	1.2	37.6	190.2	156.3	222.2	1648
1996	314.8	258.3	146.5	12.0	36.3	23.3	2.3	23.7	135.2	238.5	212.0	188.3	1591
1997	352.2	87.1	73.1	22.0	55.0	124.5	15.4	15.9	95.0	62.8	264.8	186.7	1355
1998	121.2	362.2	127.8	66.7	97.6	26.6	21.8	21.8	89.3	183.1	26.6	292.6	1437
1999	382.8	198.3	210.8	89.0	51.3	68.6	2.7	0.0	85.9	28.5	52.1	269.9	1440
2000	235.9	124.0	185.3	0.8	5.3	5.2	60.4	84.4	91.2	114.2	239.2	196.2	1342
2001	229.6	92.7	164.2	24.1	60.6	19.9	27.2	44.2	48.4	173.3	152.4	204.2	1241
2002	320.2	187.9	272.4	27.2	112.4	0.0	23.4	79.6	45.6	49.4	176.4	164.7	1459
2003	302.4	58.6	180.9	55.2	54.7	9.2	16.4	17.8	12.1	89.2	168.0	139.9	1104
2004	196.4	194.0	79.1	92.3	105.9	49.7	78.4	0.0	7.1	156.9	105.9	99.8	1166
2005	238.7	67.8	111.6	30.7	164.3	42.5	8.1	24.1	38.4	120.4	117.8	127.1	1092
2006	240.5	176.2	154.0	33.7	2.0	18.6	32.4	17.4	54.7	80.3	200.2	251.7	1262
2007	267.0	241.9	80.5	36.8	58.4	22.6	169.3	0.0	1.9	92.6	120.3	188.6	1280
2008	418.5	153.7	132.4	174.6	72.4	43.5	0.0	68.7	46.1	108.7	24.3	153.0	1396
2009	162.0	143.5	125.1	21.8	10.6	55.1	61.6	61.4	143.8	56.7	196.5	274.3	1312

a) Con esa base de datos, construya *box plots* para cada mes, y preséntelos en conjunto. Analice el comportamiento de la precipitación pluvial. Observe si se presentan valores atípicos. Además calcule las estadísticas descriptivas para cada serie mensual y discuta los resultados.

b) Además, con los totales por año, construya un gráfico de dispersión, en el eje X el tiempo y en el eje Y los valores totales anuales. Observe el comportamiento de la variación.

c) Busque una base de datos con registros de precipitación mensual de varios años, de alguna estación meteorológica cercana a su municipio de origen y realice lo citado en a) y b), en: http://www.insivumeh.gob.gt/meteorologia/boletin%20de%20estaciones%20meteorologicas.htm

16. Para estudiar el comportamiento en cuanto a su desarrollo de una planta típica de dunas, la *Hydrocotyle sp*, que es un género botánico de acuáticas o semiacuáticas formalmente clasificadas en la familia de las Apiaceae, ahora en la de las Araliaceae, se midió el tamaño del pecíolo (cm), en dos áreas: seca y húmeda. Se seleccionó de cada una de esas áreas muestras aleatorias de plantas y se midió el tamaño de los pecíolos. Los datos se presentan a continuación:

Área húmeda			
13.8	15.6	16.1	16.6
14.3	15.8	16.3	16.8
14.5	15.8	16.3	16.8
15.0	15.8	16.3	16.9
15.0	15.8	16.3	17.0
15.5	16.0	16.5	17.0
15.5	16.0	16.5	17.2
15.5	16.0	16.6	17.4
15.6	16.1	16.6	

Área seca			
7.3	8.4	9.0	10.4
7.6	8.4	9.0	10.4
7.8	8.4	9.3	10.9
7.8	8.6	9.3	10.9
8.0	8.6	9.3	11.7
8.2	8.6	9.6	11.7
8.2	8.6	9.6	12.0
8.3	9.0	9.8	
8.3	9.0	9.8	

a) Calcule la mediana, los cuartiles 1 y 3, y las distancias necesarias para poder construir un *box plot* para cada una de las áreas. Discuta los resultados obtenidos.

b) Calcule las medidas de tendencia central y de dispersión para los datos de cada área, compare y discuta los resultados.

17. Fueron tomadas dos muestras de tamaños iguales a 25 observaciones, de crecimiento de pseudobulbo, en cm, de la especie de orquídea *Laelia purpurata*, bajo condiciones de luminosidad (con luz directa y con luz indirecta). Los datos se presentan en la tabla siguiente:

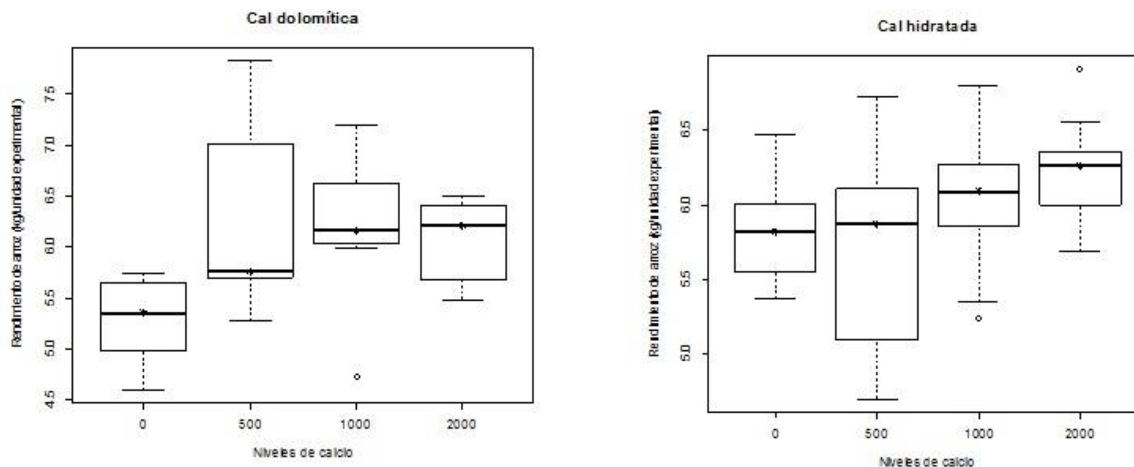
Luz directa	1.6	1.6	1.9	1.9	2.1	2.1	2.1	2.1	2.1
	2.4	2.5	2.5	2.7	3.4	3.4	3.7	3.9	4.2
	4.8	6.3	6.5	7.2	8.8	9.4	9.5		
Luz indirecta	1.4	1.9	2.8	3.1	3.5	3.5	3.6	3.9	4.3
	4.5	4.6	4.8	6.3	6.5	6.7	6.7	6.8	6.9
	8.1	8.6	10.4	12.7	16.3	16.8	16.9		

a) Calcule las medidas de tendencia central, de dispersión, asimetría y (curtosis vistas en clase) para cada conjunto de datos. Discuta los resultados obtenidos.

b) Construya un diagrama de tallos y hojas para cada conjunto de datos, indique que si cada uno de ellos tiene distribución normal.

c) Construya un *box plot* para cada conjunto de datos. Analice si las distribuciones son simétricas o no, además si existe la presencia de puntos discrepantes.

18. A continuación se presentan dos *box plots*, construidos a partir de la información proveniente de un ensayo sobre rendimientos medios de arroz (kg de grano de arroz con 14% de humedad), de la variedad ICTA Virginia, obtenidos con la aplicación al suelo de 4 niveles de Ca (0, 500, 1000 y 2000 kg.ha^{-1}) por cada fuente de Ca (cal dolomítica y cal hidratada). El experimento fue realizado en la Estación Experimental Cristina del Instituto de Ciencia y Tecnología Agrícola (ICTA), jurisdicción de la aldea Cristina, municipio de Los Amates, departamento de Izabal,



Discuta ambos gráficos con lo aprendido en la Unidad II, ¿cuáles serían sus principales conclusiones de este ensayo?

19. De una plantación de *Eucalyptus grandis*, plantada en un área de 800 ha, se desea saber cuántas parcelas de 420 m^2 deben ser muestreadas. Inicialmente fueron medidas 34 parcelas (muestra piloto), en la tabla siguiente se presentan los volúmenes (expresados en m^3/parcela) obtenidos.

Parcela	Volumen (m^3)	Parcela	Volumen (m^3)	Parcela	Volumen (m^3)
1	5.68	13	6.66	25	12.15
2	5.23	14	7.80	26	11.65
3	5.67	15	8.95	27	12.55
4	5.22	16	8.50	28	7.00
5	5.50	17	8.12	29	7.97
6	7.00	18	7.95	30	7.11
7	6.50	19	8.55	31	7.34
8	6.80	20	8.11	32	7.71
9	6.40	21	11.90	33	7.41
10	6.03	22	12.40	34	7.57
11	6.23	23	12.00		
12	5.78	24	12.23		

Calcule las medidas de tendencia central y de dispersión para datos sin agrupar. Construya un *box plot* y discuta los resultados, en términos prácticos.

20. Cinco pluviómetros fueron distribuidos en una cuenca hidrográfica de 19.4 km^2. El área de influencia de cada pluviómetro y la precipitación pluvial (en mm) reportada por cada uno de ellos, se presenta en el cuadro siguiente:

Pluviómetro	Precipitación (mm)	Área (km^2)
1	1200	1.2
2	1350	5.4
3	1412	4.9
4	1387	3.7
5	1296	4.2

Calcule la precipitación pluvial promedio de la cuenca hidrográfica.

21. Una serie familias se han clasificado por su número de hijos, resultando:

Nº de hijos	0	1	2	3	4	5	6	7	8
Nº de famílias	11	13	20	25	14	10	4	2	1

Se solicita:

a) Construya la tabla completa de frecuencias.
b) Construya las representaciones gráficas.
c) Calcule la media, mediana y moda.
d) Calcule el rango, varianza y desviación estándar.

22. Con el fin de estudiar la edad media y la dispersión de edades en un centro educativo, el director solicita estos datos a los responsables de los distintos niveles, resultando que:

200 alumnos de Primaria con media de 11 años y varianza de 2.5 u^2.
140 alumnos de Secundaria con media de 14.6 años y varianza de 2 u^2.
165 alumnos de Bachillerato con media de 17.1 años y varianza de 0.9 u^2.

¿Cuál es la edad media y la varianza del colectivo total de alumnos del centro?

Utilice las siguientes ecuaciones:

$$\overline{X} = \frac{\sum n_i . \overline{x}_i}{\sum n_i} \qquad S^2 = \frac{\sum n_i . s_i^2}{\sum n_i} + \frac{\sum n_i . \left(\overline{x}_i - \overline{X}\right)^2}{\sum n_i}$$

23. Un tren de carga realiza un trayecto de 400 km. La vía tiene trechos en mal estado que no le permitían recorrer a velocidades uniformes. Los primeros 100 km los recorre a 120km/h, los siguientes 100 km la vía está en mal estado y va a 20km/h, los terceros a 100 km/h y los 100 últimos a 130 km/h. Para calcular el promedio de velocidades, calcule la media armónica.

24. *Pinus cooperi* es un pino endémico de México, a continuación se presentan los volúmenes comerciales (en m^3) de 118 árboles de esta especie:

0.4007	0.0076	1.0940	0.3393	0.4823	0.4602
0.2993	0.3724	0.9642	0.2892	0.1639	0.3981
0.1115	0.3179	0.5686	0.1729	0.0281	0.2532
0.0189	0.1340	0.2695	0.0677	0.5313	0.1224
0.0317	0.0307	0.0564	0.4228	0.4371	0.0269
0.0264	0.1818	0.1652	0.3725	0.1619	0.6241
0.7712	0.1451	0.0762	0.1972	0.0555	0.4735
0.6707	0.0400	0.0060	0.0899	0.0094	0.2533
0.2729	0.0196	0.3844	0.0171	0.3061	0.1096
0.0611	0.5682	0.2250	0.2845	0.2183	0.4154
0.4908	0.4195	0.0449	0.1801	0.1162	0.3070
0.3005	0.1726	0.2744	0.0933	0.0401	0.1574
0.0658	0.0490	0.1911	0.0323	0.5370	0.0731
0.1385	0.3661	0.0898	0.5252	0.5228	0.0204
0.0794	0.2937	0.0284	0.4746	0.3115	0.2369
0.0162	0.1044	0.6297	0.3058	0.0938	0.2102
0.2221	0.0275	0.6125	0.1635	0.9092	0.1055
0.1950	0.1988	0.3959	0.0422	0.5930	0.0363
0.0979	0.1424	0.1910	0.9181	0.2704	
0.0383	0.0453	0.0532	0.8911	0.0892	

Pinus durangensis Ehren es un pino originario de la Sierra Madre Occidental de México. A continuación se presentan los volúmenes comerciales (en m^3) de 70 árboles de esta especie:

0.5078	0.3649	0.8246	1.7746	0.7910	0.0754	0.3218
0.1829	0.1436	0.2945	0.1833	0.1043	0.8021	0.9366
0.3428	1.0681	0.9880	0.1429	0.4254	1.5645	0.1572
0.1827	0.1670	0.6118	0.7373	0.9794	0.1859	0.1428
0.7514	0.3251	0.3509	0.2962	0.4403	0.7499	1.6231
0.1111	2.3366	0.0854	0.1936	0.1869	0.4082	0.2889
1.4673	0.1319	0.3722	1.1711	1.0576	1.0049	0.3160
0.2747	0.1328	0.3528	1.1658	0.6117	0.1473	0.2628
1.7372	0.3825	0.1646	0.1799	0.8138	1.9251	0.1082
0.1823	1.6370	0.2237	0.1122	0.0855	0.2105	0.6515

Pinus arizonica, el Pino de Arizona, es un pino mediano del norte de México, sudeste de Arizona, sudoeste de Nuevo México, y oeste de Texas en EE. UU. A continuación se presentan los volúmenes comerciales (en m^3) de 46 árboles de esta especie:

0.8948	0.1634	0.5957	0.1901	0.1619
0.4684	0.7395	0.4548	0.4478	0.9148
0.2860	0.3557	0.1192	0.5350	1.4584
0.1351	0.3110	1.4592	1.4670	1.2389
0.3873	1.0647	0.3307	1.8071	0.8515
1.8370	2.0350	0.2036	0.2197	1.6654
0.0856	0.1680	0.8999	0.1847	
2.2108	1.7064	0.7204	0.1813	
0.7602	0.3406	0.7208	2.3843	
1.6627	0.1385	0.1443	0.0934	

Para cada especie de pino realice un análisis exploratorio de datos. Compare y discuta los resultados.

25. Construya tablas (frecuencia absoluta y relativa) para resumir la información contenida en los siguientes conjuntos de datos, y calcule las medidas de tendencia central y de dispersión, así como las medidas de posición relativa.

(a) Número de plántulas en un área de 2 x 2 m: 4, 8, 15, 18, 1, 0, 17, 8, 8, 16, 8, 8, 20, 18, 4, 7, 13, 15, 1, 6, 7, 9, 3, 12, 4, 3, 7, 8, 7, 2.

(b) Altura (cm) de plantas de palo borracho (*Chorisia speciosa*) en vivero:

41.5 17.8 27.8 38.7 31.3 36.4 18.9 38.3 27.3 41.0 34.3 30.0 40.2 49.8 26.1 32.6 14.7
43.3 41.5 32.3 30.6 25.0 21.6 38.7 10.3 28.1 26.9 30.0 33.1 28.9 35.1 32.3 33.0 22.2
30.3 34.3 20.2 23.1 27.7 24.9 31.5 29.3 24.5 45.4 38.2 33.9 37.9

(c) Densidad de la madera (g/cm^3) de clones de *Eucalyptus grandis*:

0.347 0.373 0.297 0.360 0.338 0.357 0.343 0.345 0.392 0.330 0.405 0.364 0.294 0.267
0.413 0.295 0.427 0.333 0.324 0.230 0.445 0.327 0.359 0.446 0.256 0.329 0.471 0.306
0.328

26. En una empresa quieren saber la proporción media de mujeres en los diferentes departamentos. Para ello, se recolectan los datos de porcentaje de mujeres en los cinco principales departamentos, siendo estos los siguientes:

Departamento	Porcentaje
Producción	32.6
Compras	53.5
Marketing	28.9
Recursos Humanos	48.2
Administración	67.4

Como es la media de porcentajes, calcule la media geométrica que es más representativa.

27. Un profesor de Química pide a sus alumnos que realicen un experimento en el laboratorio. Espera que los alumnos obtengan 5 litros de ácido clorhídrico. Anota en una tabla una columna con las cantidades de ácido obtenidos por cada alumno y en la otra el error por falta o exceso de la cantidad esperada, de la siguiente manera:

Alumno	Litros	Error
Jorge Sánchez	5.68	0.68
Andrea Ramírez	4.62	-0.38
Raúl Prieto	3.98	-1.02
Maria José García	6.12	1.12
Amparo Sierra	5.23	0.23
Juan Pablo Rivera	4.28	-0.72

Al profesor no le importa si el error se produjo por falta o por exceso, sino la cantidad de ácido de diferencia respecto a la esperada. Para ello calcule la media cuadrática de los errores.

28. Un economista desea obtener la media del ingreso mensual por hogar (en Q), de 5 hogares, de los cuales obtuvo los siguientes resultados: Q550, Q1030, Q780, Q645 y Q975. Teniendo en cuenta que los respectivos tamaños de los hogares fueron: 2, 5, 4, 3 y 6, calcule la media ponderada.

UNIDAD III
INTRODUCCIÓN AL ESTUDIO DE PROBABILIDADES

La mayor parte de los problemas en Estadística involucran elementos de incertidumbre, ya que usualmente no es posible determinar anticipadamente las características de una población desconocida o prever las consecuencias exactas de la toma de una decisión. Por lo tanto es conveniente disponer de una medida que exprese esa incertidumbre en términos de una escala numérica. Esta medida es la PROBABILIDAD.

3.1 CONCEPTOS FUNDAMENTALES

3.1.1 EXPERIMENTO

Un experimento es el proceso mediante el cual se obtiene una observación (o una medida) de un fenómeno. Notación: ε

3.1.2 EXPERIMENTO ALEATORIO

Es el proceso de colecta de datos relativos a un fenómeno que presenta variabilidad en sus resultados.

Ejemplos de experimentos aleatorios

1. Lanzamiento de un dado y se observa el número mostrado en la cara superior.
2. Lanzamiento de una moneda cuatro veces y se observa el número de caras obtenido.
3. Una lámpara es fabricada. En seguida es probada, en cuanto a su duración, anotando el tiempo (horas) transcurrido desde que es encendida hasta que se quema.
4. Se cruzan los animales y se observa el sexo del primero que nace.
5. Se cuenta el número de larvas de gusano cogollero en plantas de maíz.
6. En una línea de producción se fabrican piezas en serie y se cuenta el número de piezas defectuosas producidas en un período de 24 horas.

NOTA: Cuando se tiene un experimento aleatorio, no se puede prever con certeza el resultado. Se puede sin embargo, describir todos los posibles resultados de ese experimento.

3.1.3 ESPACIO MUESTRAL

Es el conjunto de todos los posibles resultados de un experimento aleatorio. Notación: Ω

Ejercicio: Construya los espacios muestrales de los experimentos aleatorios del ejemplo anterior.

1. $\Omega = \{1,2,3,4,5,6\}$
2. $\Omega = \{0,1,2,3,4\}$
3. $\Omega = \{t \,/\, t \geq 0\}$
4. $\Omega = \{\text{Macho, Hembra}\}$
5. $\Omega = \{0,1,2,...\}$
6. $\Omega = \{0,1,2, \ldots, N\}$ siendo N el número máximo que puede ser producido en 24 horas.

3.1.4 EVENTO

Un evento **A** (relativo a un espacio muestral particular Ω, asociado a un experimento E) es simplemente, un conjunto de resultados posibles. En terminología de conjuntos, un evento es un subconjunto de elementos (puntos muestrales) de un espacio muestral.

Notación: **A**, B, C, D. . .

Algunos ejemplos de eventos son dados a continuación:

Nuevamente nos referiremos a los experimentos vistos anteriormente. A_i se referirá al evento asociado al experimento ε_i , i = 1, 2, 3

A_1 : Sale un número par, esto es, $A_1 = \{2, 4, 6\}$
A_2 : Ocurren dos caras, $A_2 = \{2\}$ o sea, $A_2 = \{C,C\}$
A_3 : La lámpara se quema en menos de 3 horas $A_3 = \{t \,/\, 0 \leq t < 3\}$

3.2 METODOS PARA ASIGNAR LAS PROBABILIDADES

Se pueden utilizar varios métodos para asignar las probabilidades a los resultados experimentales, sin embargo, independiente del método, se deben satisfacer dos requisitos básicos:

1. Los valores de probabilidad que se asignen a cada resultado experimental (punto muestral) deben estar comprendidos entre 0 y 1. Esto es, sí E_i representa el resultado experimental i y $P(E_i)$ representa la probabilidad de este resultado experimental, se debe cumplir:

$$0 \leq P(E_i) \leq 1 \text{ , para toda i.}$$

2. La suma de todas las probabilidades de resultados experimentales debe ser 1. Sí un espacio muestral tiene *k* resultados experimentales, se debe cumplir:

$$P(E_1) + P(E_2) + \ldots + P(E_k) = \sum_{i=1} P(E_i) = 1 \text{, o sea, } P(\Omega) = 1$$

En la práctica se usa uno de los tres métodos siguientes para asignar las probabilidades:

1. Método clásico (enfoque a priori)
2. Método de frecuencia relativa (enfoque posterior)
3. Método subjetivo

3.2.1 METODO CLÁSICO

Sí un evento A puede ocurrir en *h* maneras diferentes de un número total de *n* maneras posibles todos igualmente posibles (equiparables), entonces la probabilidad del evento es:

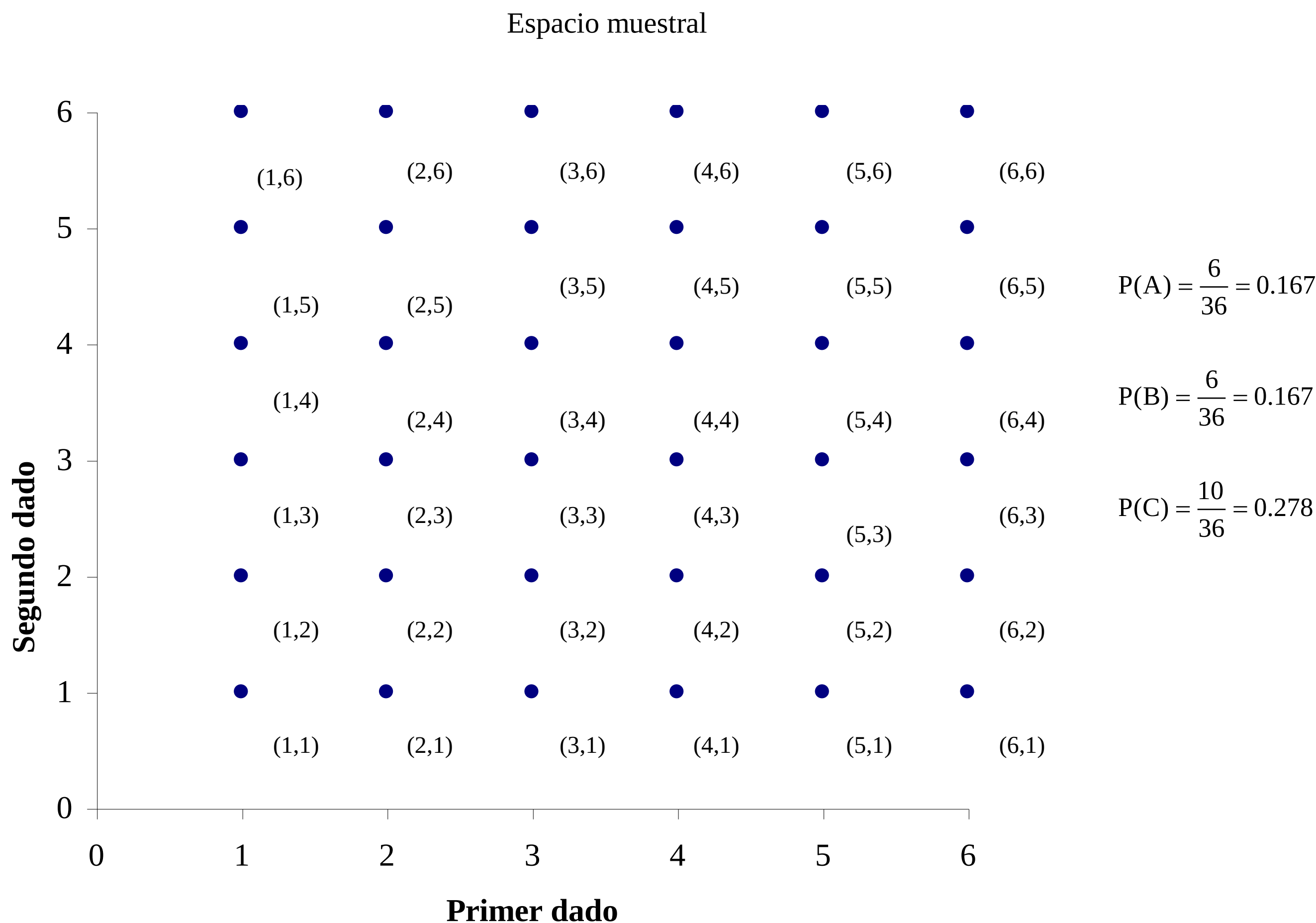
$$P(A) = \frac{h}{n} = \frac{\text{número de resultados favorables}}{\text{número de resultados probables}}$$

$\therefore$ Considerando que $A \subset \Omega$

Ejemplo 33

En el lanzamiento de dos dados honestos, calcule las probabilidades de los siguientes eventos:

A: La suma de los valores es igual a 7
B: Los resultados en los dados son iguales
C: La suma de los valores es 9 ó más.



3.2.2 METODO DE LA FRECUENCIA RELATIVA

Si después de *n* repeticiones de un experimento donde *n* es "muy grande", un evento ocurre *h* veces, entonces la probabilidad del evento es *h*/*n*.

Ejemplo 34

Si se lanza una moneda 1000 veces y se halla que 532 veces resultan caras, se puede estimar que la probabilidad de obtener una cara es 532/1000=0.532.

3.2.3 METODO SUBJETIVO

Está basado en el juicio personal. Con el método subjetivo de asignar probabilidades a los resultados experimentales, podemos usar cualquier dato disponible y también nuestra experiencia e intuición.

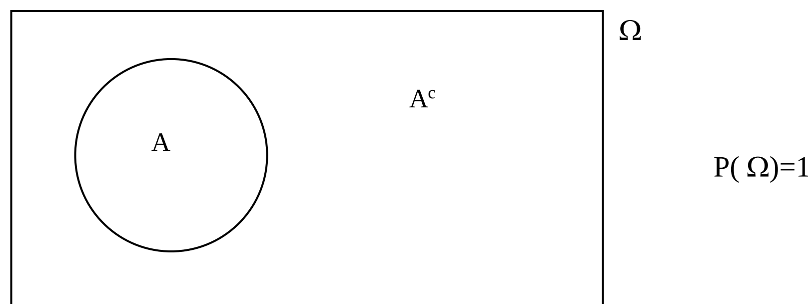
3.3 ALGUNAS RELACIONES BASICAS DE PROBABILIDAD

3.3.1 COMPLEMENTO DE UN EVENTO

Dado un evento **A**, el complemento de A se define como el evento formado por todos los puntos muestrales que no están en A, y se representa por A^c.

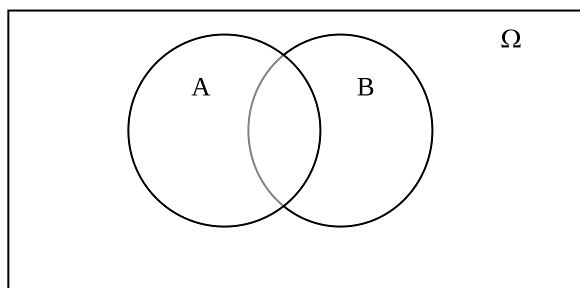
En cualquier aplicación de probabilidades, debe suceder, ya sea el evento A o su complemento A en consecuencia:

$$P(A) + P(A^c) = 1, \text{ al despejar } P(A) \text{ se obtiene } P(A) = 1 - P(A^c)$$



3.3.2 LEY ADITIVA

La ley aditiva es útil cuando se tienen dos eventos, y se desea conocer la probabilidad de que ocurra por lo menos uno de ellos. Esto es, con los eventos A y B nos interesa conocer la probabilidad de que suceda el evento A, o el evento B, o ambos. Esto es, la unión de **A** y **B** es el evento que contiene ***todos*** los puntos muestrales que pertenecen a **A** o a **B**, o a ambas, y se representa con $A \cup B$



3.3.3 INTERSECCIÓN DE DOS EVENTOS

Dados los eventos **A** y **B** la interpretación de **A** y **B** es el evento que contiene los puntos muestrales que pertenecen simultáneamente a **A** y a **B**, y se representa como **A∩B**

LEY ADITIVA: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Ejemplo 35

El gerente de personal de una empresa agroforestal encontró que el 30% de los empleados que salieron de la compañía en los dos últimos años lo hicieron principalmente por no estar satisfechos con su salario, el 20% salió por no estar satisfecho con las actividades en su trabajo y el 12% de todos los anteriores manifestaron no estar satisfechos con su salario ni con su trabajos.

¿Cuál es la probabilidad de que un empleado que haya salido en los dos últimos años lo haya hecho por no estar satisfecho con su sueldo, su trabajo o con ambas cosas?

Sean:

S = el evento de que el empleado sale debido al salario.
W = el evento de que el empleado sale por no estar satisfecho con las actividades de su trabajo.

$P(S) = 0.30$ $\quad$ $P(S \cup W) = P(S) + P(W) - P(S \cap W)$

$P(W) = 0.20$ $\quad$ $P(S \cup W) = 0.30 + 0.20 - 0.12$

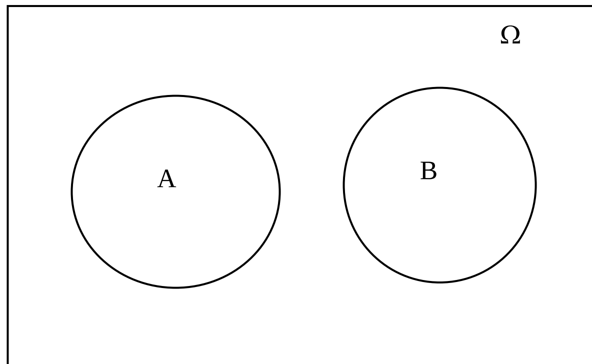
$P(S \cap W) = 0.12$ $\quad$ $= 0.38$

Respuesta:

Existe 38% de probabilidad de que un empleado salga por motivos de sueldo o de actividades de trabajo.

3.4 EVENTOS MUTUAMENTE EXCLUYENTES

Dos eventos son mutuamente excluyentes si no tienen puntos muestrales en común.



$$P(A \cup B) = P(A) + P(B) - \underbrace{P(A \cap B)}_{0}$$

$$P(A \cup B) = P(A) + P(B)$$

Ejemplo 36

Una urna contiene 12 bolas blancas y 8 negras. Si se sacan dos bolas al azar. ¿Cuál es la probabilidad de que sean del mismo color?

R/ Sean los sucesos:

A = "Extraer las dos bolas blancas"
B = "Extraer las dos bolas negras"
C = "Extraer las dos bolas del mismo color"

Según la composición de la urna se tiene que:

$$p(A) = \frac{12}{20} \bullet \frac{11}{19} = \frac{132}{380} = \frac{33}{95}$$

$$p(B) = \frac{8}{20} \bullet \frac{7}{19} = \frac{56}{380} = \frac{14}{95}$$

Como una bola no puede ser al mismo tiempo blanca y negra (los sucesos A y B son incompatibles), se tiene que:

$$p(C) = p(A) + p(B) = \frac{33}{95} + \frac{14}{95} = \frac{47}{95}$$

3.5 PROBABILIDAD CONDICIONAL

Se define como la probabilidad de un evento, dado que ha ocurrido otro evento. La probabilidad condicional de A dado B es dado por las siguientes expresiones:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad \text{siendo } P(B) > 0; \qquad P(B/A) = \frac{P(A \cap B)}{P(A)}, \quad \text{siendo } P(A) > 0$$

Ejemplo 37

Suponga que el cuadro siguiente representa la división de los alumnos del primer año de una Facultad de Agronomía, en el año de 2008. Recuerde que un alumno no puede estar matriculado en más de una carrera al mismo tiempo.

Tabla 5. Distribución de los alumnos del primer año de 1998 respecto al sexo y a la carrera.

Sexo	Carrera			Total marginal
	Agronomía (A)	Forestal (F)	Agroindustrial (E)	
Masculino (H)	160	30	15	205
Femenino (M)	40	10	10	60
Total marginal	200	40	25	265

a) Dado que un alumno es seleccionado a azar esté cursando Agronomía (A) ¿Cuál es la probabilidad de que sea del sexo masculino (H)?

$$P(H/A) = \frac{P(H \cap A)}{P(A)} = \frac{160/265}{200/265} = 0.80$$

b) Dado que el alumno seleccionado al azar es del sexo femenino (M) ¿Cuál es la probabilidad de que esté cursando Ingeniería Forestal (F)?

$$P(F/M) = \frac{P(F \cap M)}{P(M)} = \frac{10/265}{60/265} = 0.1666$$

e) ¿Cuál es la probabilidad de que un alumno seleccionado al azar este cursando Agronomía(A) dado que es de sexo femenino?

$$P(A/M) = \frac{P(A \cap M)}{P(M)} = \frac{40/265}{60/265} = 0.666$$

3.6 EVENTOS INDEPENDIENTES

Dos eventos A y B son independientes sí:

P (A/B) = P(A), o
P (B/A) = P(B).

De lo contrario, los eventos son dependientes.

3.7 LEY MULTIPLICATIVA

Mientras que la ley aditiva de la probabilidad se utiliza para determinar la probabilidad de una unión entre dos eventos, la ley multiplicativa se usa para determinar probabilidad de una intersección de dos eventos. La ley multiplicativa se basa en la definición de la probabilidad condicional.

LEY MULTIPLICATIVA: $P(A \cap B) = P(B) \times P(A/B)$, o también,

$$P(A \cap B) = P(A) \times P(B/A)$$

Ejemplo 38

El departamento de circulación de un diario sabe que el 84% de las familias de una determinada colonia tiene una suscripción para recibir el periódico de lunes a sábado. Sí D representa el evento: de que una familia tiene tal tipo de suscripción, P(D) = 0.84. Se sabe que la probabilidad de que una familia, cuya suscripción, además de ser de lunes a sábado, también se suscriba a la edición dominical (evento S), es de 0.75; esto es, P(S/D)=0.75 ¿Cuál es la probabilidad de que la suscripción de una familia incluya a la edición dominical y a la de lunes a sábado?

$$P(S \cap D) = P(D) \times P(S/D) = (0.84) \times (0.75) = 0.63$$

R/ 63% de las familias tiene una suscripción de las ediciones dominicales y entre semana.

3.8 LEY MULTIPLICATIVA PARA EVENTOS INDEPENDIENTES

$$P(A \cap B) = P(A) \times P(B)$$

La Ley Multiplicativa para eventos independientes representa otro método para determinar sí efectivamente, los eventos A y B son independientes. Esto es, sí $P(A \cap B) = P(A) \times P(B)$ entonces A y B son independientes.

Sí $P(A \cap B) \neq P(A) \times P(B)$, entonces A y B son dependientes.

Ejemplo 39

El gerente de una gasolinera sabe por su experiencia que el 80% de los clientes usan tarjeta de crédito al comprar gasolina. ¿Cuál es la probabilidad de que dos clientes consecutivos que compren gasolina usen tarjeta de crédito?

R/ A = El evento en que el primer cliente usa tarjeta de crédito
B = El evento en que el segundo cliente usa tarjeta de crédito.

$$P(A \cap B)= P(A) \times P(B)= (0.8) \times (0.8) = 0.61$$

Ejemplo 40

Se lanza un dado equilibrado (legal) 6 veces. ¿Cuál es la probabilidad de que salga algún 1 en los 6 lanzamientos?

R/ Sea el evento A="sacar algún 1 en 6 lanzamientos" y sean A1, A2, A3, A4, A5, A6, los eventos "sacar un 1 en el primero (segundo, tercero, cuarto, quinto, sexto) lanzamientos". Se tiene que:

$$p(A_1) = p(A_2) = p(A_3) = p(A_4) = p(A_5) = p(A_6) = \frac{1}{6}$$

$$p(\overline{A}_1) = p(\overline{A}_2) = p(\overline{A}_3) = p(\overline{A}_4) = p(\overline{A}_5) = p(\overline{A}_6) = \frac{5}{6}$$

Y como el evento complementario de A (no sacar ningún 1 en los seis lanzamientos) es la intersección de estos seis últimos y éstos son independientes, se tiene:

$$p(\overline{A}) = \left(\frac{5}{6}\right)^6 \text{ y } p(A) = 1 - \left(\frac{5}{6}\right)^6 = 1 - \frac{15625}{46656} = \frac{31031}{46656} = 0.665$$

Ejemplo 41

Una urna contiene dos bolas blancas y tres negras. Otra contiene seis blancas y cuatro negras. Si extraemos una bola de cada urna. ¿Cuál es la probabilidad de que sean las dos negras?
R/ Sean los eventos:
A= "sacar una bola negra de la 1ª urna"
B= "sacar una bola negra de la 2ª urna"
Se tiene que: $p(A) = \frac{3}{5}$ $\qquad p(B) = \frac{4}{10} = \frac{2}{5}$

y, dado que los dos sucesos son independientes: $p(A \cap B) = \frac{3}{5} \bullet \frac{2}{5} = \frac{6}{25}$

Ejemplo 42

Seis árboles de limón persa fueron plantados en línea recta y sabemos que dos tienen una enfermedad.

a) Si cada uno de los árboles tiene la misma susceptibilidad a estar enfermo, ¿cuál es la probabilidad de que los árboles enfermos estén a la par?

S = { (1,2) (1,3) (1,4) (1,5) (1,6) (2,3) (2,4) (2,5) (2,6) (3,4) (3,5) (3,6) (4,5) (4,6) (5,6) }

E = { los árboles enfermos estén a la par }
E = { (1,2) (2,3) (3,4) (4,5) (5,6) }

$$P(E) = \frac{5}{15} = \frac{1}{3}$$

b) Si sabemos que el árbol 3 es uno de los enfermos, ¿cuál es la probabilidad de que los árboles enfermos estén a la par?

F = { árbol 3 es uno de los enfermos }
F = { (1,3) (2,3) (3,4) (3,5) (3,6) }
$P(F) = \frac{5}{15} = \frac{1}{3}$, $E \cap F$ = { (2,3) (3,4) } y $P(E \cap F) = 2/15$.

$$P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{2/15}{5/15} = \frac{2}{5}$$

c) ¿Serán los eventos E y F independientes?

Sí: $P(E \cap F) = P(E) \times P(F)$ Los eventos son independientes.
$(2/15) \neq (5/15)(5/15)$, por lo tanto los eventos E y F no son independientes.

d) Si los árboles estuvieran plantados en círculo, y se sabe que el árbol 3 está enfermo. Calcule la probabilidad de que los árboles estén a la par. Defina si los eventos E y F son independientes.

S = { (1,2) (1,3) (1,4) (1,5) (1,6) (2,3) (2,4) (2,5) (2,6) (3,4) (3,5) (3,6) (4,5) (4,6) (5,6) }

E = {los árboles enfermos estén a la par}; E = { (1,2) (2,3) (3,4) (4,5) (5,6) (6,1) }; P(E) = 6/15

F = { árbol 3 es uno de los enfermos }; F = { (1,3) (2,3) (3,4) (3,5) (3,6) }; P (F) = 5/15
$E \cap F$ = { (2,3) (3,4) } y $P(E \cap F) = 2/15$.

Sí: $P(E \cap F) = P(E) \times P(F)$ Los eventos son independientes.
$(2/15) = (6/15)(5/15)$, por lo tanto los eventos E y F son independientes

3.9 TEOREMA DE BAYES

Este teorema también está referido como "probabilidad de las causas", es decir, probabilidad de un hecho anterior, sabiendo la probabilidad de un hecho posterior. Se basa en que los eventos definidos sobre un espacio muestral son particiones del mismo.

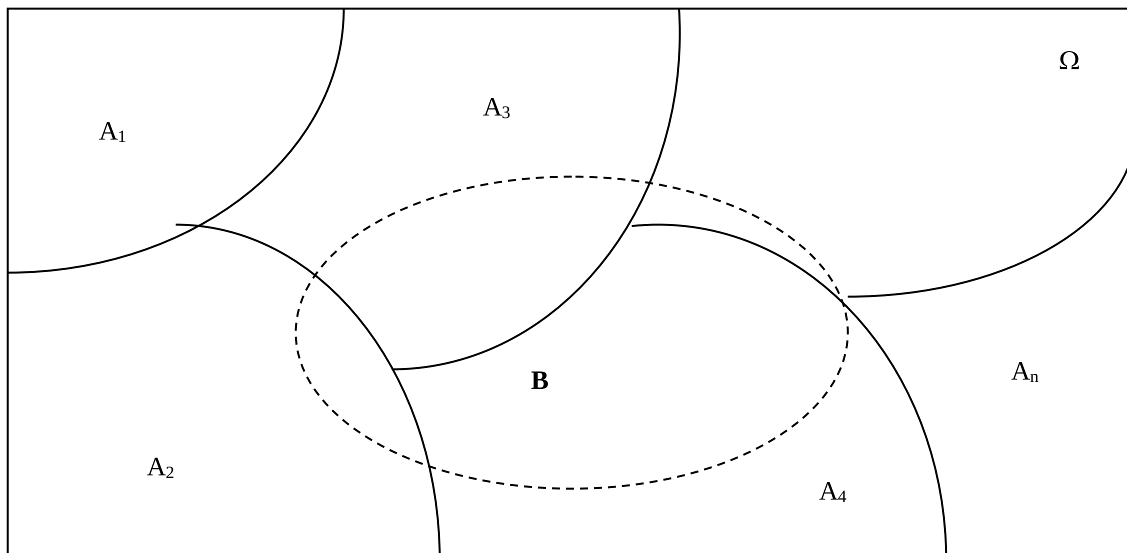
Por ejemplo, sea Ω un espacio muestral que está formado por los eventos: $A_1, A_2, A_3, \ldots, A_n$, que son mutuamente excluyentes, se tiene que:

a) $\Omega = A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n$, es decir, la unión de las particiones es igual espacio muestral.

b) No existe intersección entre las particiones,

c) La probabilidad asociada con cada una de las particiones es > 0.

Luego, si ocurre un evento B definido en Ω, se observa que:



$B \quad = \quad \Omega \cap B = (A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n,) \cap B$

$\quad = \quad (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \cup \ldots \cup (A_n \cap B)$

Donde cada uno de los eventos $A_i \cap B$ son mutuamente excluyentes, por lo que:

$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) + \ldots + P(A_n \cap B)$

Como $P(A_i)$ son las probabilidades a priori, se tiene que $P(A_i \cap B)= P(A_i) P(B/ A_i)$, o sea que la probabilidad de que ocurra el evento A_i y el evento B se obtiene a través de la multiplicación para probabilidad condicional. Por lo que se tiene:

$$P(B) = P(A_1) P(B/ A_1) + P(A_2) P(B/ A_2) + P(A_3) P(B/ A_3) + \ldots + P(A_n) P(B/ A_n)$$

Si deseamos calcular la probabilidad de que ocurra un evento A_i dado que B ya ocurrió antes, entonces:

$$P(A_i / B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i) P(B/ A_i)}{P(A_1) P(B/ A_1) + P(A_2) P(B/ A_2) + P(A_3) P(B/ A_3) + \ldots + P(A_n) P(B/ A_n)}$$

$$P(A_i / B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i) P(B/ A_i)}{\sum_{i=1}^{k} P(A_i) P(B/ A_i)}$$

La expresión anterior es el teorema de Bayes, que como se observa, es una simple probabilidad condicional.

Ejemplo 43

Una fábrica con 3 sucursales producen 40, 35 y 25% del total de la producción. Tiene los siguientes porcentajes de artículos defectuosos: 4, 6 y 8%, respectivamente. Sí se elige aleatoriamente un artículo, calcule las siguientes probabilidades:

a) de que no sea defectuoso. R/ 0.9430
b) si resultó defectuoso, ¿cuál es la probabilidad de que proceda de la primera sucursal? $P(A_1/B)= 0.2807$
c) si no resultó ser defectuoso ¿Cuál es la probabilidad de que proceda de la segunda sucursal? $P(A_2/C)= 0.3489$

Solución:

A_1 = el producto es de la primera sucursal
A_2 = el producto es de la segunda sucursal
A_3 = el producto es de la tercera sucursal
B = el producto es defectuoso
C = el producto no es defectuoso

Eventos A_i	Probabilidades previas $P(A_i)$	Probabilidades condicionales $P(B/A_i)$	Probabilidades conjuntas $P(A_i \cap B) = P(A_i)\, P(B/A_i)$	Probabilidades posteriores $P(A_i /B)$
A_1	0.40	0.04	0.016	0.2807
A_2	0.35	0.06	0.021	0.3684
A_3	0.25	0.08	0.020	0.3509
Σ	1.00		P(B)= 0.057	1.00

$$P(B/A_i) = \frac{P(A_i \cap B)}{P(A_i)}, \quad \text{siendo } P(A_i) > 0 \qquad\qquad P(A_i/B) = \frac{P(A_i \cap B)}{P(B)}, \quad \text{siendo } P(B) > 0$$

Eventos A_i	Probabilidades previas $P(A_i)$	Probabilidades condicionales $P(C/A_i)$	Probabilidades conjuntas $P(A_i \cap C) = P(A_i)\, P(C/A_i)$	Probabilidades posteriores $P(A_i /C)$
A_1	0.40	0.96	0.3840	0.4072
A_2	0.35	0.94	0.3290	0.3489
A_3	0.25	0.92	0.2300	0.2439
Σ	1.00		P(C) = 0.943	1.0000

3.10 PRINCIPIO FUNDAMENTAL DEL CONTEO

❑ PRINCIPIO MULTIPLICATIVO

Sí un primer suceso (algunos autores lo citan como evento) puede efectuarse de P_1 maneras diferentes, y si después de que este suceso ha sido efectuado, un segundo suceso puede efectuarse de P_2 maneras diferentes, . . . , y finalmente un *k*- ésimo suceso puede realizarse en P_k maneras diferentes, entonces todos los k sucesos pueden realizarse en el orden especificado en $P_1 \times P_2 \times \ldots \times P_k$ maneras diferentes.

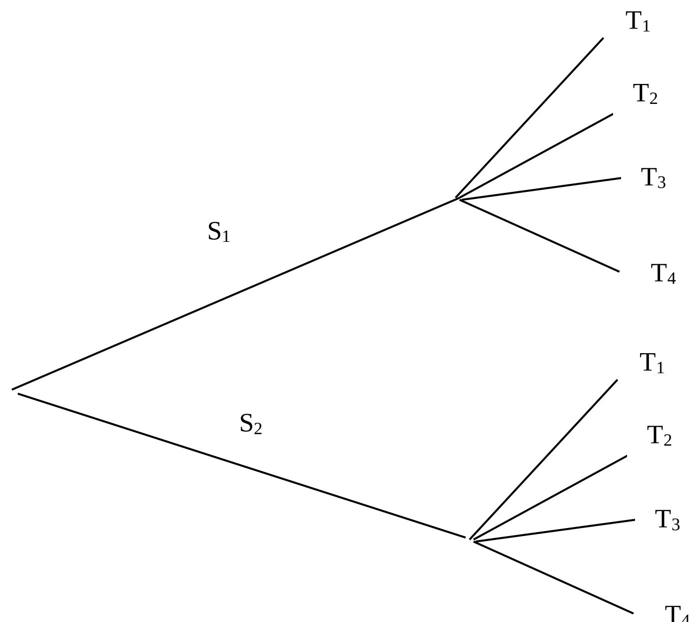
Ejemplo 44

Sí un hombre tiene 2 camisas y 4 corbatas entonces tiene $2 \times 4 = 8$ maneras de escoger una camisa y luego una corbata.

❑ DIAGRAMA DE ARBOL

Un diagrama, llamado diagrama de árbol (debido a su apariencia), se emplea frecuentemente en conexión con el principio anterior.

Solución: Sí las camisas se representan por S_1, S_2 y las corbatas por T_1, T_2, T_3 y T_4, las diferentes maneras de escoger una camisa y luego una corbata se indican en el diagrama de árbol siguiente:



FACTORIAL

El producto de cualquier número entero positivo *n* por todos los enteros menores que *n* se llama factorial de *n* y se expresa por el símbolo *n*! Por lo tanto:

0! = 1 por definición
1! = 1 (1) = 1
2! = 2 (1) = 2
3! = 3 (2) (1) = 6
4! = 4 (3) (2) (1) = 24
.
.
.

$n = (n)\ (n-1)\ (n-2), \ldots, (1)$

El factorial de los primeros números enteros positivos se pueden obtener directamente utilizando una calculadora común, para números mayores se obtienen con la ecuación aproximada de Stirling: $n! \sim \sqrt{2\,\pi\,n}\; n^n e^{-n}$, siendo *e*=2.71828 . . . la base de los logaritmos naturales.

El símbolo ~ en la ecuación de Stirling indica que la relación del lado izquierdo al lado derecho se aproxima a 1 a medida que $n \rightarrow \infty$.

Ejemplo 45: Hallar el valor 50!

$50! \sim \sqrt{2\,\pi\,50}\; 50^{50} e^{-50} \equiv N$, para evaluar N se utilizan logaritmos de base 10. Así:

$$\log N = \log\left(\sqrt{100\,\pi}\;50^{50}e^{-50}\right)$$

$$= \frac{1}{2}\log 100 + \frac{1}{2}\log \pi + 50\log(50) - 50\log e$$

$$= \frac{1}{2}\log(2) + \frac{1}{2}(0.4972) + 50\,(1.6990) - 50\,(0.4343)$$

log N = 64.4836

N = antilog (64.4836) = 3.04×10^{64}

PERMUTACIONES

Supóngase que se tienen *n* objetos diferentes y deseamos ordenar *r* de estos objetos en una línea. Puesto que hay *n* maneras de escoger el primer objeto, y luego de hacer esto *n*–1 maneras de escoger el segundo objeto, . . , y finalmente *n* – r +1 formas de escoger el *r*–ésimo objeto, se deduce por el principio fundamental del conteo que el número de ordenaciones, o permutaciones diferentes como generalmente se les llama, está dado por:

$$n\,Pr = n\,(n-1)\,(n-2)\ldots(n-r-1)$$

nPr = número de permutaciones de *n* objetos tomados de *r* en *r*.

Para el caso particular cuando *r* = *n* , nPr se convierte en: $n\,Pr = n\,(n-1)\,(n-2)\ldots 1 = n!$, que se denomina *n* factorial. En términos factoriales nPr se puede escribir como:

$$n\,Pr = \frac{n!}{(n-r)!}$$

Ejemplo 46: Calcule el número de permutaciones diferentes que pueden tomarse con las letras A,B,C,D,E,F,G, tomando 3 a la vez.

$${}_7P_3 = \frac{7!}{(7-3)!} = \frac{7!}{4!} = \frac{7\times 6\times 5\times 4!}{4!} = 210$$

COMBINACIONES

En una permutación interesa el orden de la distribución de los objetos. Así abc es una permutación diferente de bca. Sin embargo, en muchos problemas interesa solamente seleccionar o escoger objetos sin interesar su orden. Dichas selecciones se llaman: COMBINACIONES. Por ejemplo abc y bca son la misma combinación.

El número total de combinaciones de *r* objetos seleccionados de *n* se denota por nCr ó $\binom{n}{r}$ y está dado por:

$$\binom{n}{r} = nCr = \frac{n!}{r!(n-r)!}$$

Ejemplo 47: ¿De cuántas formas puede elegirse una comisión de 5 personas de entre 9 personas?

$$\binom{9}{5} = {}_9C_5 = \frac{9!}{5!(9-5)!} = \frac{9!}{5!\ 4!} = \frac{9\times8\times7\times6\times5\times4!}{5!\ 4!} = 126$$

Ejemplo 48

Se escriben a azar las cinco vocales. ¿Cuál es la probabilidad de que la “e” aparezca la primera y la “o” la última?

R/ Al escribir al azar las 5 vocales tenemos P5= 5! = 120 casos posibles. De entre ellos, si la e ha de aparecer la primera y la o la última, tenemos las otras 3 vocales que han de permutar en los tres lugares centrales, es decir, los casos favorables son P3= 3!=6. La probabilidad pedida es:

$$p = \frac{6}{120} = \frac{1}{20}$$

LISTA DE EJERCICIOS 3

1. Suponga que 3% de una población de adultos ha intentado suicidarse. También se sabe que 20% de esa población vive en condiciones de extrema pobreza. Sí estos dos eventos son independientes, ¿cuál es la probabilidad de que un individuo seleccionado al azar haya intentado suicidarse y además viva en condiciones de extrema pobreza?

2. En un taller hay 3 máquinas; la primera se avería al mes con una probabilidad de 0.04, la segunda con 0.06 y la tercera con 0.1; sus averías son independientes en probabilidad. Se pide:

   a) Probabilidad de que se averíe una sola máquina en el mes;
   b) Probabilidad de que se averíen las tres máquinas en el mes;
   c) Probabilidad de que se averíen la primera y la segunda, pero no la tercera.

3. El Sr. Fernández está dudando entre dedicar sus ahorros a un viaje a Cuba o invertir en renta variable. Su asesor fiscal le ofrece dos alternativas atrayentes, pero él ante su falta de formación bursátil, confía al azar su decisión. Invertirá en el sector eléctrico si saca una bola roja de una urna que contiene 20 bolas, de las cuales 8 son rojas, 3 verdes y 9 negras. Si la bola no es roja lanzará dos dados y si obtiene una suma de 6 entre ambos invertirá en el sector inmobiliario; en caso contrario se decidirá por las vacaciones en Cuba. ¿Cuál es la probabilidad de que finalmente disfrute del viaje?

4. Una población está clasificada en tres grupos, según la edad: el 20% está entre 25 y 35 años, el 65% entre 36 y 50 años y el 15% entre 51 y 65 años. Al investigar los hábitos de dicha población se ha comprobado que toman café por la mañana el 70% del grupo del primer grupo de edades, el 40% del segundo y el 10% del tercero.

   a) Seleccionado aleatoriamente un individuo de la población ¿cuál es la probabilidad de que sea del grupo de 25 a 35 años y tome café?

   b) Si sabemos que un individuo toma café ¿cuál es la probabilidad de que pertenezca al grupo de 51 a 65 años?

5. El 40 % de los alumnos de la Facultad de Ciencias Económicas y Empresariales de la UNED proceden de otra Universidad, el 25 % estudia su segunda carrera y el resto cursa estudios superiores por primera vez. El porcentaje de mujeres en cada uno de estos grupos es de 40, 60 y 55 respectivamente. Para elaborar una encuesta se elige al azar un estudiante y se desea saber:

   a) Cual es la probabilidad de que proceda de otra Universidad y sea mujer.
   b) Si se eligió una mujer, ¿cuál es la probabilidad de que proceda de otra Universidad?

6. Por un estudio encargado por el partido político "Seguro Que Ganamos" (SQG) se obtiene la siguiente información: el 17% de la población tiene estudios superiores, el 44% estudios medios, el 30% estudios primarios y el 9% no tiene estudios. De entre los de estudios superiores el 25% votan al partido SQG, entre los de estudios medios el 35%, entre los de estudios primarios el 22% y entre los que no tienen estudios votan partido SQG el 18%. Si se extrae un sujeto al azar, obtenga las siguientes probabilidades:

   a) Que sea titulado superior sabiendo que vota al SQG.
   b) Que sea persona sin estudios que vota al SQG;
   c) Que sea una persona con estudios primarios o que no vote al SQG.

7. Suponga que un espacio muestral es: S = {E_1, E_2, E_3, E_4, E_5, E_6, E_7} en donde de E_1, . . . ,E_7 representan los puntos muestrales. Se aplican las siguientes asignaciones de probabilidades: $P(E_1)=0.05$, $P(E_2)=0.20$, $P(E_3)=0.20$, $P(E_4)=0.25$, $P(E_5)=0.15$, $P(E_6)=0.10$, $P(E_7)=0.05$

Sean $A=\{E_1, E_4, E_6\}$ a) Determinar P(A), P(B), P(C)

$B=\{E_2, E_4, E_7\}$ b) Determinar $P(A\cup B)$

$C=\{E_2, E_3, E_5, E_7\}$ c) Determinar $P(A\cap B)$

d) ¿Son mutuamente excluyentes A y C? e) Determinar B^c y P(BC)

8. Suponga que una caja contenga diez bolas distribuidas de la siguiente manera:

- Tres son de color y tienen puntos
- Una es de color y tienen franjas
- Dos son grises y tienen puntos
- Cuatro son grises y tienen franjas

a) Si una persona extrae de la caja una bola de color:

a.1 ¿Cuál es la probabilidad de que ésta contenga puntos?
a.2 ¿Cuál es la probabilidad de que tenga franjas?

b) ¿Cuál es la probabilidad de la bola tenga puntos, dado que es gris?
c) Calcule P (gris/puntos) y P (color/puntos).

9. Una tienda de autoservicio ha sido víctima de muchos ladrones durante un mes determinado, pero debido al aumento de las condiciones de seguridad de la tienda, se ha podido aprehender a 250 ladrones. Se registró el sexo de cada infractor y si éste era su primero robo o si ya había sido sorprendido con anterioridad. Los datos se resumen en la siguiente tabla:

Sexo	Primera aprehensión	Segunda aprehensión
Hombre	60	70
Mujer	44	76

Suponiendo que un infractor aprehendido es seleccionado al azar, encuentre:

a) La probabilidad de que éste sea hombre.
b) La probabilidad de que sea la primera aprehensión del infractor, dado que sea hombre.
c) La probabilidad de que el ladrón sea mujer, dado que éste es reincidente.
d) La probabilidad de que el ladrón sea mujer, dado que es su primera aprehensión.

10. El gerente regional de una compañía privada de paquetería está preocupada por la posibilidad de que algunos de sus empleados vayan a huelga. Estima que la probabilidad de que sus pilotos para viajes fuera de la ciudad (P1) vayan a huelga es de 0.75 y la probabilidad de que sus pilotos para viajes dentro de la ciudad (P2) se vayan a huelga es de 0.65. Además estima que si los P2 se van a huelga, existe 90% de posibilidad de que los P1 realicen un paro solidario de actividades.

a) ¿Cuál es la probabilidad de que ambos grupos se vayan a huelga?
b) Si los pilotos P1 hacen huelga ¿Cuál es la probabilidad de que los pilotos P2 lo hagan también como acto de solidaridad?

11. Un transportista de productos tiene 10,000 cajas de plátanos que provienen de Ecuador y de Honduras. Una inspección de la carga ha dado la siguiente información:

Origen	Número de cajas con fruta:		Total de cajas
	Echada a perder	Muy madura	
Ecuador	200	840	6000
Honduras	365	295	4000

a) ¿Cuál es la probabilidad de que una caja seleccionada al azar contenga fruta echada a perder?

b) ¿Cuál es la probabilidad de que una caja seleccionada al azar contenga fruta muy madura?

c) Dado que una caja seleccionada al azar contiene fruta muy madura ¿Cuál es la probabilidad de que provenga de Honduras?

12. En un conjunto de estudiantes el 15% estudia alemán, el 30% estudia francés y el 10% ambas materias.

a) ¿Son independientes los sucesos estudiar alemán y estudiar francés?

b) Si se elige un estudiante al azar, calcule la probabilidad de que no estudie francés ni alemán.

13. En una ciudad el 55% de los habitantes consume pan integral, el 30% consume pan de multicereales y el 20% consume ambos. Se pide:

a) Sabiendo que un habitante consume pan integral, ¿cuál es la probabilidad de que coma pan de multicereales?

b) Sabiendo que un habitante consume pan de multicereales, ¿cuál es la probabilidad de que no consuma pan integral?

c) ¿Cuál es la probabilidad de que una persona de esa ciudad no consuma ninguno de los dos tipos de pan?

14. La probabilidad de que un hombre viva más de 25 años es de 3/5, la de una mujer es de 2/3. Se pide:

a) La probabilidad de que ambos vivan más de 25 años.
b) La probabilidad de que sólo viva más de 25 años el hombre.
c) La probabilidad de que sólo viva más de 25 años la mujer.
d) La probabilidad de que viva más de 25 años, al menos, uno de los dos.

15. La probabilidad de que un artículo provenga de una fábrica A_1 es 0.7, y la probabilidad de que provenga de otra A_2 es 0.3. Se sabe que la fábrica A_1 produce 4 por mil de artículos defectuosos y la A_2 8 por mil.

a) Se observa un artículo y se ve que está defectuoso. ¿Cuál es la probabilidad de que provenga de la fábrica A_2?
b) Se pide un artículo a una de las dos fábricas, elegida al azar. ¿Cuál es la probabilidad de que esté defectuoso?
c) Se piden 5 artículos a la fábrica A_1 ¿Cuál es la probabilidad de que haya alguno defectuoso?

16. De las 100 personas que asisten a un congreso 40 hablan francés, 40 inglés, 51 castellano, 11 francés e inglés, 12 francés y castellano y 13 inglés y castellano. Se eligen al azar dos asistentes y se desea saber:

a) ¿Cuál es la probabilidad de que ninguno hable francés?
b) ¿Cuál es la probabilidad de que hablen castellano?
c) ¿Cuál es la probabilidad de que se entiendan sólo en castellano?
d) ¿Cuál es la probabilidad de que sólo hablen un idioma?
e) ¿Cuál es la probabilidad de que hablen los tres idiomas?

17. En un hospital especializado en enfermedades de tórax ingresan un 50 % de enfermos de bronquitis, un 30 % de neumonía y un 20 % con gripe. La probabilidad de curación completa en cada una de dichas enfermedades es, respectivamente, 0.7; 0.8 y 0.9. Un enfermo internado en el hospital ha sido dado de alta completamente curado. Hallar la probabilidad de que el enfermo dado de alta hubiera ingresado con bronquitis.

18. Una empresa productora de papel y celulosa dispone de 250 registros de candidatos para algunas plazas vacantes. Se asume que los registros representan una muestra aleatoria de la población económicamente activa de la ciudad. En los registros, 60% son hombres y 40% son mujeres. Se sabe que en esta ciudad el 50% de los hombres son fumadores, y apenas 20% de las mujeres fuman.

a) ¿Cuál es la proporción de la población que calificaría para un empleo de motosierrista (hombre y que no fume)?

b) ¿Cuál es la probabilidad de seleccionar al azar una persona de sexo femenino, si se sabe que no es fumadora?

19. En los parcelamientos “La Máquina” “Nueva Concepción” y “La Blanca” se producen en su orden el 14, 18, y 25% del total de maíz de la república. Los porcentajes de grano podrido son en su orden: 3,4 y 6%. Para el resto del país, el porcentaje de grano podrido es el triple de la desviación estándar de los porcentajes de pudrición anteriores.

Sí se escoge un saco al azar y se descubre que está podrido:

a) ¿Cuál es la probabilidad de que provenga de “La Máquina”?
b) ¿Cuál es la probabilidad de que provenga del resto del país?
c) ¿Cuál es la probabilidad de que siendo de “La Blanca” esté podrido?
d) Si se escoge otro y no está podrido ¿Cuál es la probabilidad de que provenga de la Nueva Concepción?

20. Pablo y Pedro juegan 12 partidas de ajedrez, de las cuales Pablo gana 6, Pedro 4 y terminan 2 empatadas. Acuerdan jugar un torneo consistente en 3 partidas. Hallar la probabilidad de que:

a) Pablo gane las 3 partidas. R/ 1/8
b) Dos partidas terminen empatadas. R/ 5/72
c) Pablo y Pedro ganen alternativamente. R/ 5/36
d) Pedro gane al menos una partida. R/ 19/27

21. Un botiquín contiene 2 frascos de aspirinas y 3 de tabletas para la gripe. Un segundo botiquín contiene 3 de aspirinas, 2 de tabletas para la gripe y 1 de tabletas laxantes. Sí se toma un frasco aleatoriamente de cada botiquín, encuentre la probabilidad de que:

a) ambos frascos contengan tabletas para la gripe,
b) ningún frasco contenga tabletas para la gripe;
c) los dos frascos contengan diferentes tabletas.

22. Entre 60 partes de refacción automotriz cargadas en un camión en Guatemala, 45 tienen a Quetzaltenango por destino y 15 a Huehuetenango. Si dos de las partes se descargan por error en Escuintla y la "selección" es aleatoria, ¿qué probabilidades hay de que:

a) ambas partes debieran de haber llegado a Quetzaltenango,
b) ambas partes debieran de haber llegado a Huehuetenango,
c) una debiera haber llegado a Quetzaltenango y la otra a Huehuetenango. R/ a.33/59 b. 7/118 c.45/118

23. La probabilidad de que un Ingeniero Agrónomo diagnostique correctamente una enfermedad en un cultivo en particular es de 0.7. Dado que realizó un diagnóstico incorrecto, la probabilidad de que el encargado del cultivo levante una demanda es de 0.9. ¿Cuál es la probabilidad de que el Ingeniero Agrónomo realice un diagnóstico incorrecto y de que el encargado lo demande? R/0.27

24. "Si el 80% de la población adulta ve televisión y el 70% lee algún periódico, demuestre que por lo menos el 50% acude a ambos medios de comunicación."

25. En un grupo de inválidos de guerra, 70% ha perdido un ojo, 75% una oreja, 80% un brazo y 85% una pierna. ¿Cuál es la probabilidad mínima que hayan perdido los cuatro miembros? ¿Y la máxima?

26. Un grupo de excursionistas está realizando un paseo por el parque de Las Victorias de Cobán (Alta Verapaz), en un momento dado se encuentran con tres posibles caminos, a los que llamaremos A, B y C. La posibilidad de que tomen cualquier camino es la misma. Se sabe que la probabilidad de que realicen la ruta sin perderse si toman el camino A es 0.7, si toman el B 0.8 y el C 0.9. Si se sabe que han acabado la ruta y no se han perdido ¿Cuál es la probabilidad de que hayan tomado el camino B?

27. En una población de bovinos se presenta una epidemia. El 10 % de los machos y el 18 % de las hembras están enfermos. Se sabe además que hay doble número de hembras que de machos y se pide:

a) Elegido al azar un individuo de esa población ¿Cuál es la probabilidad de que esté enfermo?
b) Un individuo de esa población se sabe que está enfermo ¿Qué probabilidad hay de que el citado individuo sea macho?

28. Se hace una encuesta en un grupo de 120 personas, preguntando si les gusta leer y ver la televisión. Los resultados son:

A 32 personas les gusta leer y ver la tv.
A 92 personas les gusta leer.
A 47 personas les gusta ver la tv

Si elegimos al azar una de esas personas:

¿Cuál es la probabilidad de que no le guste ver la tv?
¿Cuál es la probabilidad de que le guste leer, sabiendo que le gusta ver la tv?
¿Cuál es la probabilidad de que le guste leer?

29. El 1% de la población de un determinado lugar padece una enfermedad. Para detectar esta enfermedad se realiza una prueba de diagnóstico. Esta prueba da positiva en el 97% de los pacientes que padecen la enfermedad; en el 98% de los individuos que no la padecen da negativa. Si elegimos al azar un individuo de esa población:

a) ¿Cuál es la probabilidad de que el individuo dé positivo y padezca la enfermedad?
b) Si sabemos que ha dado positiva, ¿cuál es la probabilidad de que padezca la enfermedad?

30. En una clase de 30 alumnos hay 18 que han aprobado matemáticas, 16 que han aprobado inglés y 6 que no han aprobado ninguna de las dos. Elegimos al azar un alumno de esa clase:

a) ¿Cuál es la probabilidad de que haya aprobado inglés y matemáticas?
b) Sabiendo que ha aprobado matemáticas, ¿cuál es la probabilidad de que haya aprobado inglés?
c) ¿Son independientes los sucesos "Aprobar matemáticas" y "Aprobar inglés"?

31. Un estudiante cuenta, para un examen con la ayuda de un despertador, el cual consigue despertarlo en un 80% de los casos. Si oye el despertador, la probabilidad de que realiza el examen es 0.9 y, en caso contrario, de 0.5.

a) Si va a realizar el examen, ¿cuál es la probabilidad de que haya oído el despertador?

b) Si no realiza el examen, ¿cuál es la probabilidad de que no haya oído el despertador?

32. En una casa hay tres llaveros A, B y C; el primero con cinco llaves, el segundo con siete y el tercero con ocho, de las que sólo una de cada llavero abre la puerta de un gabinete de cocina. Se escoge al azar un llavero y, de él, una llave intenta abrir el gabinete. Se pide:

a) ¿Cuál será la probabilidad de que se acierte con la llave?

b) ¿Cuál será la probabilidad de que el llavero escogido sea el tercero y la llave no abra?

c) Y si la llave escogida es la correcta, ¿cuál será la probabilidad de que pertenezca al primer llavero A?

33. En un bosque de *Pinus elliottii*, 30% de los árboles fueron resinados. De los árboles no resinados, 70% son apropiados para aserrío, en tanto que dentro de los resinados apenas 10% lo son. Asumiendo que en un árbol de este bosque es seleccionado al azar, calcule:

a) ¿Cuál es la probabilidad de ser apropiado para aserrío?

b) ¿Cuál es la probabilidad de que el árbol haya sido resinado y ser propio para aserrío?

c) ¿Cuál es la probabilidad de que haya sido resinado y no sea propio para aserrío?

d) ¿Cuál es la probabilidade de que no haya sido resinado y no sea propio para aserrío?

34. Un alumno de Ingeniería Forestal considera las oportunidades de conseguir dos centros de práctica profesional. Las probabilidades de conseguir la práctica en una empresa forestal son de 80%, en tanto que las probabilidades de conseguir una práctica en un parque nacional son de 70%. Las probabilidades de conseguir en ambos lugares son de 50%.

Calcule:

a) ¿Cuál es la probabilidad que el alumno consiga la práctica en el parque nacional, dado que él consiguió la práctica en una empresa?

b) ¿Cuál es la probabilidad que el alumno consiga la práctica en una empresa forestal, dado que él consiguió la práctica en el parque nacional?

c) ¿Cuál es la probabilidad que el aluno consiga realizar la práctica en por lo menos uno de los lugares?

3.11 VARIABLES ALEATORIAS

Muchos experimentos producen resultados no numéricos. Antes de analizarlos es conveniente transformar sus resultados en números, lo que es realizado a través de la **VARIABLE ALEATORIA** (o variable estocástica)**,** que es una regla de asociación de un valor numérico a cada punto del espacio muestral. Comúnmente las variables aleatorias se denotan por una letra mayúscula (X,Y,Z, por ejemplo)

DEFINICIÓN: Una variable X es una variable aleatoria, sí los valores que toma X y que corresponden a los diferentes resultados de un experimento, son eventos fortuitos o aleatorios.

Una variable aleatoria puede ser de uno de dos tipos, discreto o continuo. Sí el número de valores que puede tomar la variable aleatoria es enumerable entonces se le llama: **VARIABLE ALEATORIA DISCRETA.** Por otra parte sí una variable aleatoria puede tomar o asumir cualquier valor dentro de un intervalo dado, entonces se trata de una **VARIABLE ALEATORIA CONTINUA.**

3.11.1 DISTRIBUCIÓN DE PROBABILIDAD DE UNA VARIABLE ALEATORIA DISCRETA

Sea X una variable aleatoria discreta y suponiendo que los valores posibles que puede tomar están dados por x_1, x_2, x_3, . . ., dispuestos en orden creciente de magnitud. Suponiendo también que los valores se asumen con probabilidades dadas por: $P(X=x_i) = f(x_i) = p_i$, i=1,2, . . ., la distribución de probabilidad de una variable aleatoria discreta es un tabla que representa el conjunto de los valores de la variable y sus respectivas probabilidades de ocurrencia obtenidos a través de una función $f(x_i)$.

X	.x_1	x_2	x_3	.	.	.
p_i	.p_1	p_2	p_3	.	.	.

Observación:

- $0 \leq p_i \leq 1$ y $\sum_{i=1} p_i = 1$
- X=x, representa el evento "la variable aleatoria X toma el valor x", y p(X=x) representa la probabilidad de dicho evento.

Ejemplo 49

Consideremos el experimento: lanzamiento de dos monedas legales (no viciadas) y sea X el número de caras observadas en el lanzamiento.

El espacio muestral asociado a ese experimento es dado por:

Ω={ (cara, cara); (cara, escudo); (escudo, cara); (escudo, escudo) }

Sí X es el número de caras y Y el número de escudo, tenemos que, para cada punto muestra podemos asignar un número:

Posibles Resultados	X	Y
(cara,cara)	2	0
(cara,escudo)	1	1
(escudo, cara)	1	1
(escudo, escudo)	0	2

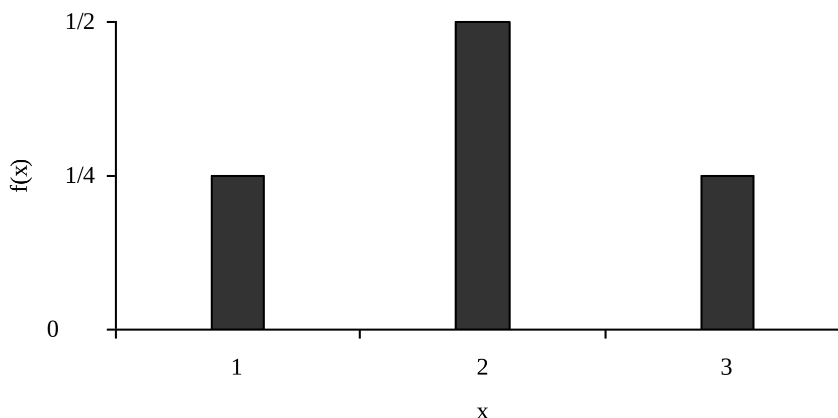
Con esta información, se puede encontrar la función correspondiente a la variable aleatoria X, P(CC)= ¼, P(CE)= ¼, P(EC)= ¼, P(EE)= ¼.

P(X=0) = P(EE) = ¼
P(X=1) = P(CE) + P(EC) = ¼ + ¼ = ½
P(X=2) = P(CC) = ¼

Entonces, la distribución de probabilidad es dada por:

X	0	1	2
f(x)	¼	½	¼

Esta distribución de probabilidad puede ser representada a través de la gráfica de probabilidad:



3.11.2 FUNCION DE DISTRIBUCIÓN PARA UNA VARIABLE ALEATORIA DISCRETA

En muchas ocasiones no interesa tanto conocer la probabilidad de que la v.a. X tome exactamente un determinado valor x_i, cuanto la probabilidad de que tome valores menores o iguales que un cierto valor x_i. En tales casos es necesario acumular los distintos valores de la función de probabilidad hasta el valor deseado. Se trata de una nueva aplicación llamada *función de distribución*.

La función de distribución acumulada, o simplemente, la función de distribución para una variable aleatoria discreta X, se define por:

$$P(X \leq x) = F(x).$$

Siendo x cualquier número real, es decir, $-\infty < x < \infty$.

Si X únicamente toma un número finito de valores $x_1, x_2, x_3, \ldots, x_n$ entonces la función de distribución está dada por:

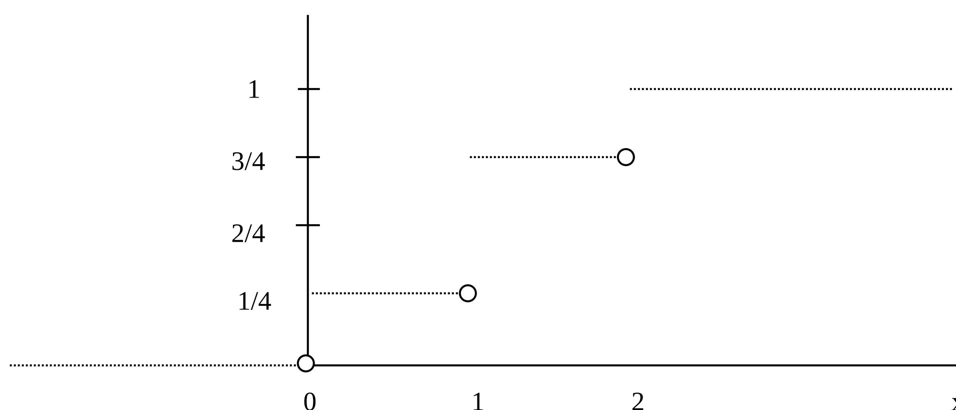
$$F(x) = \begin{cases} 0 & -\infty < x < x_1 \\ f(x_1) & x_1 \le x < x_2 \\ f(x_1) + f(x_2) & x_2 \le x < x_3 \\ \vdots & \vdots \\ f(x_1) + f(x_2) + \ldots f(x_n) & x_n \le x < \infty \end{cases}$$

Ejemplo 50

a) Encuentre la función de distribución para el ejemplo del lanzamiento de la moneda.

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1/4 & 0 \le x < 1 \\ 3/4 & 1 \le x < 2 \\ 1 & 2 \le x < \infty \end{cases}$$

b) Represente gráficamente la función distribución de probabilidad



Debido a la apariencia de la anterior gráfica, frecuentemente se le llama: Función escalera.

Nota:
$(X<x)$ representa el suceso "la variable aleatoria X toma un valor menor a x", y
$p(X<x)$ representa la probabilidad de que la v.a. X tome un valor menor a x.
$(X \le x)$ representa el suceso "la variable aleatoria X toma un valor menor o igual a x", y
$p(X \le x)$ representa la probabilidad de que la v.a. X tome un valor menor o igual a x.

LISTA DE EJERCICIOS 4

1. Una variable aleatoria X tiene la siguiente función de distribución de probabilidad:

$$F(x) = \begin{cases} 0 & -\infty < x < 10 \\ 0.2 & 10 \leq x < 12 \\ 0.5 & 12 \leq x < 13 \\ 0.9 & 13 \leq x < 25 \\ 1 & 25 \leq x < \infty, \end{cases}$$

Con esta información:

a) Construya la gráfica de la función de distribución de probabilidad
b) Encuentre la función de probabilidad f(x)
c) Calcule:

c.1) $P(X \leq 12)$,
c.2) $P(X<12)$,
c.3) $P(12 \leq X \leq 20)$ y
c.4) $P(X \geq 18)$

2. Una empresa distribuye sierras en exclusiva en Guatemala y las recibe semanalmente de la fábrica ubicada en Alemania. El número X de miles de sierras vendidas cada mes, es una variable aleatoria cuya función de densidad es:

$$f(x) = \begin{cases} k(1-x)^3, \text{ si } 0 < x < 1 \\ 0, \text{ caso contrario} \end{cases}$$

Se pide:

a) Hallar el valor de k.
b) Calcular el promedio de unidades vendidas al mes.
c) Si el distribuidor quiere tener una garantía del 95 % de que no se le agote el producto en un mes determinado, ¿qué cantidad del mismo debe tener almacenado?

3.11.3 MEDIDAS DE POSICIÓN PARA VARIABLES ALEATORIAS DISCRETAS: VALOR MEDIO O ESPERANZA MATEMÁTICA

Definición:

Dada una variable aleatoria X, con valores posibles x_1, x_2, x_3 , . . . , x_n y valores de probabilidad $p(x_i) = P(X=x_i)$, i=1,2, . . ., k, entonces, el valor esperado de X denotado por E(X) es definido como:

$$E(X) = \sum_{i=1}^{k} x_i p(x_i)$$

La esperanza matemática frecuentemente se le conoce como la media de X, y suele denotarse también por μ_X.

Propiedades de la esperanza matemática (en general, para variables aleatorias discretas y contínuas)

1. Sí X = C, siendo C una constante, entonces E(X)=C

2. Suponga que C sea una constante y X una variable aleatoria, entonces E(C*X)= C*E(X)

3. Sean X y Y dos variables aleatorias cualquiera, entonces E(X+Y)=E(X) + E(Y)

4. Si X y Y son dos variables aleatorias cualquiera, entonces $E(XY) = E(X) \times E(Y)$

Ejemplo 51

Considere una variable aleatoria discreta X con función de probabilidad discreta dada por:

X	2	5	8	15	20
p_i	0.1	0.3	0.2	0.2	0.2

Encuentre el valor esperado de X:

$E(X) = (2\times0.1) + (5\times0.3) + (8\times0.2) + (15\times0.2) + (20\times0.2) = 10.3$ (es el centro de la distribución de probabilidad).

3.11.4 MEDIDAS DE DISPERSION PARA VARIABLES ALEATORIAS DISCRETAS: VARIANZA

Sea X una variable aleatoria discreta con valores de probabilidad $p(x_i) = P(X=x_i)$, i=1,2, . . ., k, y media μ_X, la varianza de X denotada por V(X) se define de la siguiente manera:

$$V(X) = E\left[(X-\mu_X)^2\right] = E\left[X - E(X)\right]^2$$

El cálculo de V(X) puede ser simplificado con el auxilio del siguiente resultado:

$$V(X) = E(X^2) - \left[E(X)\right]^2$$

Demostración:

$$V(X) = E\left[X - E(X)\right]^2 = E\left\{X^2 - 2X \times E(X) + \left[E(X)\right]^2\right\}$$

Considerando que E(X) es una constante, E[E(X)]=E(X), entonces:

$$\begin{aligned} V(X) &= E(X^2) - 2E(X) \times E(X) + \left[E(X)\right]^2 \\ &= E(X^2) - 2\left[E(X)\right]^2 + \left[E(X)\right]^2 \\ &= E(X^2) - \left[E(X)\right]^2 \end{aligned}$$

¿Qué mide la varianza? Mide la dispersión de la variable alrededor de la media.

Ejemplo 52

En un cierto barrio de la ciudad de México, las compañías de seguros establecieron el siguiente modelo para el número de vehículos robados por semana:

X	0	1	2	3	4
p_i	¼	½	1/8	1/16	1/16

Calcule la media y la varianza del número de robos semanales en ese barrio:

$E(X) = (0\times1/4) + (1\times1/2) + (2\times1/8) + (3\times1/16) + (4\times1/16) = 1.188 \approx 1.19$ robos

$E(X^2) = (0^2\times1/4) + (1^2\times1/2) + (2^2\times1/8) + (3^2\times1/16) + (4^2\times1/16) = 2.563$

$Var(X) = 2.563 - (1.19^2) = 1.151\ u^2$

Propiedades de la varianza de una variable aleatoria

1. Sí C es una constante, $V(X + C) = V(X)$
2. Sí C es una constante, $V(CX) = C^2 V(X)$
3. Sí (X,Y) son variables independientes, entonces: $V(X + Y) = V(X) + V(Y)$
4. Sí (X,Y) son variables dependientes, entonces: $V(X + Y) = V(X) + V(Y) + 2\,Cov(X,Y)$

Demostración:

$$\begin{aligned} V(X + Y) &= E\left[(X + Y) - (\mu_X + \mu_Y) \right]^2 \\ &= E\left[(X - \mu_X) - (Y - \mu_Y) \right]^2 \\ &= E\left[(X - \mu_X)^2 + 2(X - \mu_X)(Y - \mu_Y) + (Y - \mu_Y)^2 \right] \\ &= E(X - \mu_X)^2 + 2E\left[(X - \mu_X)(Y - \mu_Y)\right] + E(Y - \mu_Y)^2 \\ &= V(X) + V(Y) + 2\,Cov(X, Y) \end{aligned}$$

3.11.5 MEDIDA DE ASOCIACION ENTRE DOS VARIABLES ALEATORIAS: COVARIANZA

La covarianza, denotada por Cov (X,Y), es el valor esperado del producto de los desvíos de cada variable con relación a su media, y está dada por la siguiente expresión:

$$\text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

A partir de esta expresión, se puede llegar a otra expresión más simple:

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

Demostración:

$$\begin{aligned} \text{Cov}(X,Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - X\mu_Y - Y\mu_X + \mu_X \mu_Y] \\ &= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X \mu_Y \end{aligned}$$

Recordemos que $E(X) = \mu_X$ y que $E(Y) = \mu_Y$, entonces:

$$\begin{aligned} \text{Cov}(X,Y) &= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - 2E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

Ejemplo 53

Una región de la Costa Sur de Guatemala fue dividida en 10 subregiones. En cada una de ellas, fueron observadas dos variables: número de pozos artesanales (X) y el número de riachuelos o ríos presentes en la subregión (Y). Los resultados son presentados en la tabla siguiente:

Subregión	1	2	3	4	5	6	7	8	9	10
X	0	0	0	0	1	2	1	2	2	0
Y	1	2	1	0	1	0	0	1	2	2

Considerando que se escoge una de las subregiones al azar, esto es, cada subregión tiene la misma probabilidad 1/10 de ser seleccionada, podemos construir la distribución conjunta de (X,Y):

(X,Y)	Probabilidad
(0,0)	1/10
(0,1)	2/10
(0,2)	2/10
(1,0)	1/10
(1,1)	1/10
(2,0)	1/10
(2,1)	1/10
(2,2)	1/10
Total	1

Una forma equivalente de presentar la distribución conjunta es a través de la tabla de doble entrada:

X \ Y	0	1	2	P(X=x)
0	1/10	2/10	2/10	5/10
1	1/10	1/10	0	2/10
2	1/10	1/10	1/10	3/10
P(Y=y)	3/10	4/10	3/10	1

Por tanto, las funciones de probabilidad marginales son las siguientes:

X	0	1	2
p_i	5/10	2/10	3/10

Y	0	1	2
p_i	3/10	4/10	3/10

Si se conoce la distribución conjunta de X y Y, el comportamiento de otras variables, tales como X+Y o XY puede ser determinado, como se ilustra a continuación:

(X,Y)	X+Y	XY	$P(x,y)$
(0,0)	0	0	1/10
(0,1)	1	0	2/10
(0,2)	2	0	2/10
(1,0)	1	0	1/10
(1,1)	2	1	1/10
(2,0)	2	0	1/10
(2,1)	3	2	1/10
(2,2)	4	4	1/10

Utilizando la tabla anterior, la función de probabilidad de X+Y y la de XY son obtenidas fácilmente:

X+Y	0	1	2	3	4
Probabilidad	1/10	3/10	4/10	1/10	1/10

XY	0	1	2	4
Probabilidad	7/10	1/10	1/10	1/10

Para los valores esperados se tiene:

$$E(X+Y) = (0\times 1/10) + (1\times 3/10) + (2\times 4/10) + (3\times 1/10) + (4\times 1/10)$$
$$= 18/10$$

$$E(XY) = (0\times 7/10) + (1\times 1/10) + (2\times 1/10) + (4\times 1/10)$$
$$= 7/10$$

$$E(X) = (0\times 5/10) + (1\times 2/10) + (2\times 3/10)$$
$$= 8/10$$

$E(Y) = (0\times 3/10) + (1\times 4/10) + (2\times 3/10)$
$= 1$

Como $E(XY) = 7/10 \neq E(X)\,E(Y) = (8/10)\,(1)$, las variables aleatorias X y Y no son independientes, por tanto $Cov(X,Y) \neq 0$.

$Cov(X,Y) = E(XY) - E(X)\,E(Y)$
$= 7/10 - (8/10)\,(1)$
$= -1/10$

$V(X + Y) = V(X) + V(Y) + 2\,Cov(X,Y)$
$V(X) = E(X^2) - [E(X)]^2 = 14/10 - (8/10)^2 = 76/100$
$E(X^2) = (0^2\times 5/10) + (1^2\times 2/10) + (2^2\times 3/10)$
$= 14/10$
$V(Y) = E(Y^2) - [E(Y)]^2 = 16/10 - (1)^2 = 60/100$
$E(Y^2) = (0^2\times 3/10) + (1^2\times 4/10) + (2^2\times 3/10)$
$= 16/10$

$V(X + Y) = 76/100 + 60/100 + 2\,Cov\,(-1/10) = 116/100$

3.12 DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

3.12.1 DISTRIBUCIÓN BINOMIAL

Un experimento binomial es un experimento que posee las siguientes características:

1. Consta de n ensayos o pruebas idénticas (ensayos de Bernoulli)
2. Cada prueba puede tener uno de dos resultados posibles (éxito o fracaso)
3. La probabilidad de un éxito en una sola prueba es igual a *p*, y permanece constante de una a otra prueba. En tanto, la probabilidad de fracaso es igual a (1-*p*) y se denota con la letra *q*.
4. El resultado obtenido en cada prueba es independiente de los resultados obtenidos anteriormente.

La distribución Binomial se suele representar por *B(n,p)* siendo *n* y *p* los parámetros de dicha distribución.

La probabilidad de obtener x éxitos está dada por la siguiente función de probabilidad:

$$P(X = x) = nCx\ p^x\ q^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

$$nCx = \frac{n!}{x!(n-x)!}, \text{ siendo que:}$$

n = número de pruebas o ensayos.
p = probabilidad de éxito en una sola prueba.
q = probabilidad de fracaso (1-*p*)

Parámetros de la Distribución Binomial

Media = E(X) = n p
Varianza = V(X) = npq

Ejemplo 54

Calcule la probabilidad de que en una familia de 4 hijos por lo menos uno sea niño. Considere la variable X número de niños varones. Suponga que la probabilidad del nacimiento de un niño es ½.
Solución:

$P(X=1) = {}_4C_1\ (1/2)^1\ (1/2)^{4-1} \quad = \dfrac{4!}{4!(4-1)!}(1/2)^1\ (1/2)^3 = 1/4$

$P(X=2) = {}_4C_2\ (1/2)^2\ (1/2)^{4-2} \quad = \dfrac{4!}{4!(4-2)!}(1/2)^2\ (1/2)^2 = 3/8$

$P(X=3) = {}_4C_3\ (1/2)^3\ (1/2)^{4-3} \quad = \dfrac{4!}{4!(4-3)!}(1/2)^3\ (1/2)^1 = 1/4$

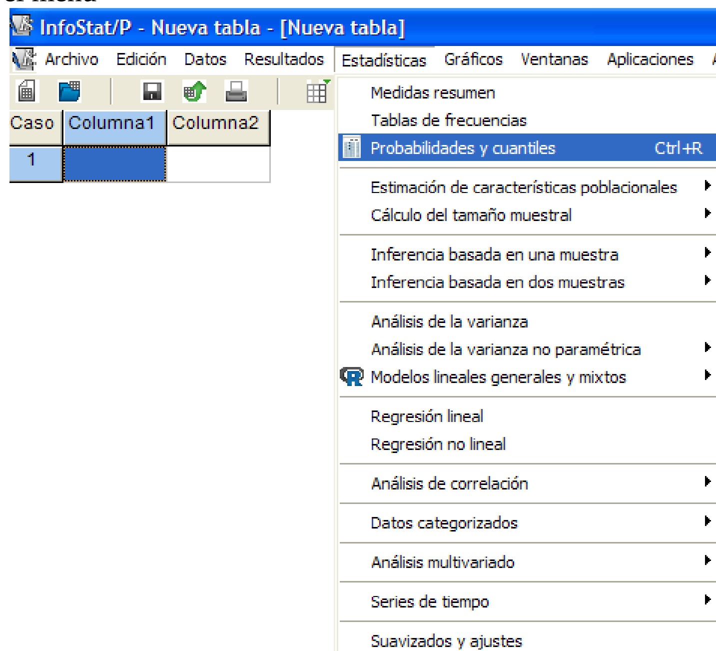
$P(X=4) = {}_4C_4\ (1/2)^4\ (1/2)^{4-4} \quad = \dfrac{4!}{4!(4-4)!}(1/2)^4\ (1/2)^0 = 1/16$

$P(X \geq 1) = \sum_{i=1}^{4} P(x_i) = P(X=1)+P(X=2)+P(X=3)+P(X=4) = 15/16$

Otra forma: $P(X \geq 1) = 1-P(X=0) = 1-1/16 = 15/16$

Resolución del Ejemplo 54 con Infostat

1. Ingreso desde el menú



2. Verificar resultados y comparar con los cálculos manuales

Calculador de probabilidades y cuantiles

Seleccione la distribución
Uniforme (a,b)
Normal (media,varianza)
T Student (v)
Chi Cuadrado (v,lambda)
F no central (u,v,lambda)
Exponencial (lambda)
Gamma (lambda,r)
Beta (a,b)
Weibull (a,b)
Doble Exponencial (a,b)
Logística (a,b)
Pareto (Theta,a)
Gumbel (a,b)
Rangos estudent.(k,v)
Poisson (lambda)
Binomial (n,p)
Geométrica (p)
Hipergeométrica (m,k,n)
Binomial Negativa (m,k)
Beta-Binomial(p,Rho,N)

4 n
0.5 p
Valor de x
1
Prob. (X<=x)
0.3125
Prob. (X>x)
0.6875
Prob. (X=x)
0.25
Calcular
Ayuda

Ejemplo 55: De 2000 familias con 4 niños ¿Cuántas calcula que deben tener:

a) Al menos un niño

(2000) (15/16) = 1875 familias

b) Exactamente 2 niños

(2000) (3/8) = 750 familias

c) Ninguna niña

(2000) (1/16) = 125 familias

Ejemplo 56: Un examen de selección múltiple, del curso de Ecología Vegetal consta de 100 preguntas, cada una de ellas con 6 posibles respuestas (de las cuales solamente una es correcta).

a) ¿Cuál será la calificación esperada para una persona que no tiene conocimiento del material de la prueba?

b) Entre qué limites caerán las calificaciones de no conocimiento?

Solución:
n = 100 preguntas
p = 1/6
q = 5/6

$E(X) = n\,p$
$= 100\,(1/6) = 16.7 \approx 17$ puntos

Para encontrar la variación de las calificaciones de no conocimiento, se necesita saber el valor de desviación estándar de la variable:

$$\sigma_X = \sqrt{npq} = \sqrt{(100)(1/6)(5/6)} = 3.7 \text{ puntos}$$

Por el teorema de Tchebychev se esperaría que el 95% de las calificaciones de no conocimiento estén en el intervalo $\mu_X \pm 2\,\sigma_X$, o sea, 16.7 ±(2)(3.7), o sea, de 9.3 hasta 24.1 puntos.

3.12.2 DISTRIBUCIÓN DE POISSON

La distribución de Poisson se llama así en honor a Simeón Dennis Poisson (1781-1840), francés que desarrolló esta distribución. La distribución de Poisson se puede utilizar también para aproximar una distribución de probabilidad binomial cuando n es “grande” y p es “pequeño”, y cuando E(X) = n p de la distribución binomial es aproximadamente menor que 7.

La distribución de Poisson es un buen modelo para la distribución de frecuencias relativas del número de eventos raros que ocurren en una unidad de tiempo, de distancia de espacio, etc.

Una variable aleatoria X tiene distribución de Poisson con parámetro λ >0, si su función de probabilidad es dada por:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0,1,2,\ldots,\infty$$

λ = lambda, representa el número medio de ocurrencias por intervalo de tiempo.
e = 2.71728 (base de los logaritmos neperianos o naturales)

La notación utilizada será: $X \sim P_o(\lambda)$

Parámetros de la Distribución de Poisson

Media	= E(X) =	λ	=	n p
Varianza	= V(X) =	λ		

Ejemplo 57

Supóngase que se está investigando la seguridad de un crucero muy peligroso. Los archivos de la policía indican una media de cinco accidentes por mes en él. El número de accidentes está distribuido conforme a la distribución de Poisson, y la división de seguridad en carreteras quiere calcular la probabilidad de ocurrencia de exactamente 0,1,2,3 y 4 accidentes en un mes determinado.

Aplicando el modelo de Poisson, se tiene que:

$P(0) = (5)^0\,(e^{-5})\,/0! = 0.00674$

$P(1) = (5)^1\,(e^{-5})\,/1! = 0.03370$

$P(2) = (5)^2 (e^{-5}) /2! = 0.08425$

$P(3) = (5)^3 (e^{-5}) /3! = 0.14042$

$P(4) = (5)^4 (e^{-5}) /4! = 0.17552$

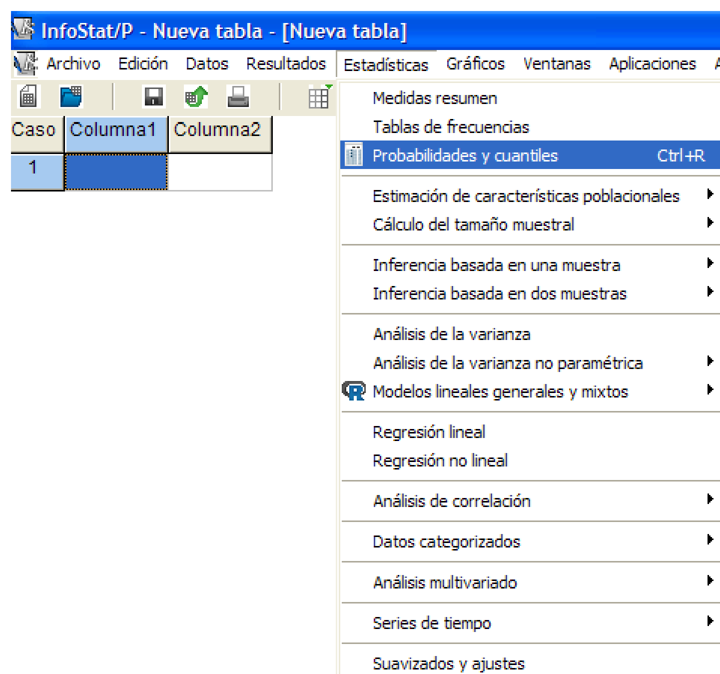
Para saber cuál es la probabilidad de 3 ó menos, se suman las probabilidades de 0,1,2,3 lo que será igual a :

P(0) = 0.00674
P(1) = 0.03370
P(2) = 0.08425
P(3) = 0.14042
P(3 o menos) = 0.26511

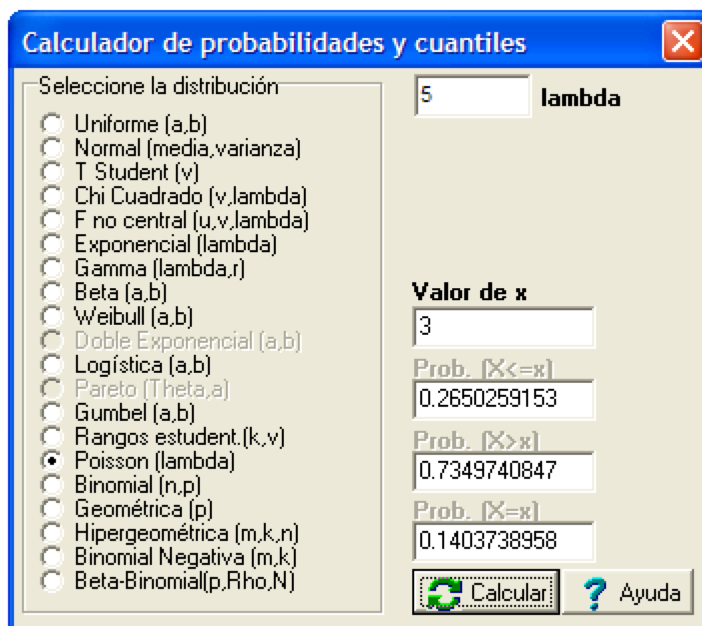
Dado que la probabilidad de que haya 3 o menos accidentes es de 0.26511 entonces la probabilidad de que ocurran más de tres debe ser = 1 –0.26511 = 0.73489.

Resolución del Ejemplo 57 con Infostat

i. Ingreso desde el menú



ii. Verificar resultados y comparar con los cálculos manuales



Ejemplo 58

Si la probabilidad de que una persona sufra una reacción dañina al ingerir un determinado antibiótico es de 0.001. Calcule la probabilidad de que de un total de 3000 pacientes sufran el malestar:

a) Exactamente 3 personas

b) Más de 3 personas presenten reacción dañina

Solución:

$\lambda = (3000)\ (0.001) = 3$

a) $$P(X=3)=\frac{e^{-3}3^{3}}{3!}=0.2240$$

b) $$P(X>2) = 1-\sum_{i=0}^{2}P(x_i)$$

P(X=0) = 0.0498		P(X>2) = 1 – 0.4232 =	0.5768
P(X=1) = 0.1494			
P(X=2) = 0.2240			
0.4232			

Ejemplo 59: Aproximación de la distribución Poisson a la distribución binomial

Se puede utilizar la distribución de probabilidad de Poisson como aproximación a la distribución binomial cuando:

a) La probabilidad de éxito p es pequeña ($p \leq 0.05$), y

b) El número n de ensayos es grande ($n \geq 20$)

c) Sí $n \geq 100$, la aproximación es generalmente excelente, siempre y cuando $np \leq 10$.

Por ejemplo: si se sabe que el 5% de los libros encuadernados en cierto taller tienen encuadernaciones defectuosas. Determine la probabilidad de que 2 de 100 libros encuadernados en ese taller, tengan encuadernaciones defectuosas, usando, a) la distribución binomial, b) la aproximación de Poisson a la distribución binomial.

Solución:

a) n = 100
p = 0.05 = p(encuadernación defectuosa) = p(éxito)
q = 0.95 = p(encuadernación no defectuosa) = p(fracaso)
x = variable que define el número de encuadernaciones defectuosas en la muestra= 0, 1, 2, 3,....,100 encuadernaciones defectuosas

$$P(x = 2, n = 100, p = 0.05) = {}_{100}C_2(0.05)^2(0.95)^{98} = (4950)(0.05)^2(0.95)^{98} = 0.0812$$

b) n = 100 encuadernaciones
p = 0.05
$\lambda = np = (100)(0.05) = 5$

x = variable que define el número de encuadernaciones defectuosas en la muestra = 0, 1, 2, 3,....,100 encuadernaciones defectuosas

$$p(x = 2, \lambda = 5) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{5^5(2.718)^{-5}}{2!} = 0.0843$$

Al comparar los resultados de las probabilidades con una y otra distribución, se observa que la diferencia entre un cálculo y otro es de tan solo 0.0031, por lo que la aproximación de Poisson es una buena opción para calcular probabilidades binomiales.

3.12.3 DISTRIBUCIÓN GEOMÉTRICA O DE PASCAL

La distribución geométrica es un modelo adecuado para aquellos procesos en los que se repiten pruebas hasta la consecución del éxito o resultado deseado y tiene interesantes aplicaciones en los muestreos realizados de esta manera. También implica la existencia de una dicotomía de posibles resultados y la independencia de las pruebas entre sí.

Esta distribución se puede hacer derivar de un proceso de Bernoulli en el que se tienen las siguientes características:

a) El proceso consta de un número no definido de pruebas o experimentos separados o separables. El proceso concluirá cuando se obtenga por primera vez el resultado deseado (éxito).

b) Cada prueba puede dar dos resultados mutuamente excluyentes: k y no k

c) La probabilidad de obtener un resultado k en cada prueba es p y la de obtener un resultado no k es q siendo (p + q = 1).

d) Las probabilidades p y q son constantes en todas las pruebas, por tanto, las pruebas son independientes (si se trata de un proceso de "extracción" éste se llevará a cabo con devolución del individuo extraído).

$$X \sim G(p)$$

$$P(X=k) = q^{k-1} * p$$

Propiedades:

Esperanza: $E(X) = 1/p$
Varianza: $V(X) = q/p^2$

Ejemplo 60

Si el 25% de la población del DF está a favor del candidato Roberto Madrazo para las elecciones del 2006.

a) Encuentre la probabilidad que la primera persona que esté a favor del candidato Madrazo, se encuentre después de la quinta persona entrevistada.

b) ¿Cuántas personas se espera entrevistar hasta encontrar la primera que esté a favor del candidato Madrazo?

Solución:

X = cantidad de personas que se van a entrevistar aleatoriamente hasta obtener la primera que esté a favor del candidato Roberto Madrazo.

P = 0.25 y q = 0.75

a) $P(x > 5) = 1 - P(x \leq 5) = 1 - 0.7627 = 0.2373$

$P(x = 1) = 0.75^{1-1} * 0.25 = 0.25$
$P(x = 2) = 0.75^{2-1} * 0.25 = 0.1875$
$P(x = 3) = 0.75^{3-1} * 0.25 = 0.1406$
$P(x = 4) = 0.75^{4-1} * 0.25 = 0.1055$
$P(x = 5) = 0.75^{5-1} * 0.25 = 0.0791$

$P(x > 5) = 1 - P(x \leq 5) = 1 - 0.7627 = 0.2373$

Usando Infostat: $P(X=k) = q^k * p$

Algunos autores consideran la aleatorización como "número de pruebas anteriores al primer éxito". De esta manera el conseguir el éxito a la primera sería X=0, de tal manera que: **$P(X=k) = q^k * p$.**

Esperanza: $E(X) = (q/p) + 1$
Varianza: $V(X) = q/p^2$

$P(x > 4) = 1 - P(x \leq 4) = 1 - 0.7627 = 0.2373$

$P(x = 0) = 0.75^0 * 0.25 = 0.25$
$P(x = 1) = 0.75^1 * 0.25 = 0.1875$
$P(x = 2) = 0.75^2 * 0.25 = 0.1406$
$P(x = 3) = 0.75^3 * 0.25 = 0.1055$
$P(x = 4) = 0.75^4 * 0.25 = 0.0791$

b) Esperanza: $E(X) = 1/p = 1/0.25 = 4$
Usando Infostat $E(X) = (q/p) + 1 = (0.75/0.25) + 1 = 4$

3.12.4 DISTRIBUCIÓN BINOMIAL NEGATIVA

Sí x es igual al número de fracasos antes de obtener k éxitos, entonces la variable aleatoria X tiene función de probabilidad dada por la ecuación:

$$P(X = x) = \binom{k + x - 1}{x} p^k q^x$$

Siendo:
p = probabilidad de éxito
q = probabilidad de fracaso
k = cantidad de éxitos
x = cantidad de fracasos

Ejemplo 61

La probabilidad que un alumno que no entienda binomial negativa repruebe el examen es de 75% si se pide seleccionar 5 alumnos reprobados al azar. Calcular la probabilidad de haber tomado 3 alumnos aprobados antes de los 5 reprobados.

p = 0.75   q = 0.25
k = 5   x = 3

Solución:

Argumentos de función
NEGBINOMDIST
Núm_fracasos 3 = 3
Núm_éxitos 5 = 5
Prob_éxito 0.75 = 0.75
= 0.129776001
Devuelve la distribución binomial negativa, la probabilidad de encontrar núm_fracasos antes que núm_éxito, con probabilidad_s de un suceso.
Prob_éxito es la probabilidad de éxito; un número entre 0 y 1.
Resultado de la fórmula = 0.129776001
Ayuda sobre esta función
Aceptar Cancelar

Esperanza: $E(X) = k(1-p)/p = 5(0.25/0.75) = 1.66666...$

En Infostat:

Calculador de probabilidades y cuantiles
Seleccione la distribución
Uniforme (a,b)
Normal (media,varianza)
T Student (v)
Chi Cuadrado (v,lambda)
F no central (u,v,lambda)
Exponencial (lambda)
Gama (lambda,r)
Beta (a,b)
Weibull (a,b)
Doble Exponencial (a,b)
Logística (a,b)
Pareto (Theta,a)
Gumbel (a,b)
Rangos estudent.(k,v)
Poisson (lambda)
Binomial (n,p)
Geométrica (p)
Hipergeométrica (m,k,n)
Binomial Negativa (m,k)
Beta-Binomial(p,Rho,N)
1.6666 media
5 k
Valor de x
3
Prob. (X<=x)
0.8861950744
Prob. (X>x)
0.1138049256
Prob. (X=x)
0.1297708098
Calcular
Ayuda

3.12.5 DISTRIBUCIÓN HIPERGEOMÉTRICA

La función de probabilidad para esta función está dada por la ecuación:

$$P(X=x) = h(x,n,M.N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Siendo:

N = tamaño de la población
n = tamaño de la muestra
M = número de éxitos en la población
x = número de éxitos en la muestra

Ejemplo 62

Una caja contiene 20 bombones, de los cuales 8 son de caramelo y 12 de nueces. Si una persona selecciona al azar 4 bombones. ¿Cuál es la probabilidad de seleccionar:

a) 1 de caramelo?
b) 2 de caramelo?

c) 3 de caramelo?
d) 4 de caramelo?
e) Ninguno de caramelo?

Para P(x=1) en Excel:

Argumentos de función
DISTR.HIPERGEOM
Muestra_éxito 1 = 1
Núm_de_muestra 4 = 4
Población_éxito 8 = 8
Núm_de_población 20 = 20
= 0.363261094
Devuelve la probabilidad para una variable aleatoria discreta siguiendo una distribución hipergeométrica.
Núm_de_población es el tamaño de la población.
Resultado de la fórmula = 0.363261094
Ayuda sobre esta función
Aceptar Cancelar

En Infostat:

Calculador de probabilidades y cuantiles
Seleccione la distribución
Uniforme (a,b)
Normal (media,varianza)
T Student (v)
Chi Cuadrado (v,lambda)
F no central (u,v,lambda)
Exponencial (lambda)
Gama (lambda,r)
Beta (a,b)
Weibull (a,b)
Doble Exponencial (a,b)
Logística (a,b)
Pareto (Theta,a)
Gumbel (a,b)
Rangos estudent.(k,v)
Poisson (lambda)
Binomial (n,p)
Geométrica (p)
Hipergeométrica (m,k,n)
Binomial Negativa (m,k)
Beta-Binomial(p,Rho,N)
20 m
8 k
4 n
Valor de x
1
Prob. (X<=x)
0.4654282766
Prob. (X>x)
0.5345717234
Prob. (X=x)
0.3632610939
Calcular Ayuda

LISTA DE EJERCICIOS 5

1. Si X ~ B (15,0.4), encontrar los siguientes valores de probabilidad:
   a) P (X ≥ 14)
   b) P (8< X ≤ 10)
   c) P (X < 2)

2. Un equipo de fútbol tiene probabilidad de victoria igual a 0.92 siempre que juega. Si el equipo juega 4 partidos, determine la probabilidad de que gane:

   a) Todos los juegos.
   b) Exactamente 2 juegos.
   c) Por lo menos un juego.
   d) A lo sumo 3 juegos.

3. En una planta industrial los lotes grandes de artículos recibidos se inspeccionan para detectar los defectuosos, por medio de un esquema de muestreo. Se examinan 10 artículos, el lote será rechazado si se encuentran 2 o más artículos defectuosos. Si un lote contiene 5% de artículos defectuosos. ¿Cuál es la probabilidad de que:

   a) El lote sea aceptado.
   b) El lote sea rechazado.

4. Un experimento consiste en la siembra de 50 semillas de maíz híbrido, las cuales tienen 85% de poder germinativo. Con base en la anterior información, calcule los siguientes valores de probabilidad:

   a) Que germinen 8 semillas.
   b) Por lo menos 3 semillas germinen.
   c) Calcule la esperanza matemática y la varianza.

5. Sea X ~ B (n,p), y sabiendo que E(X) = 12 y Var (X) = 3, determine:
   a) n
   b) p
   c) P (X ≥ 14)
   d) P (X < 12)

6. Calcule la probabilidad de acertar correctamente por lo menos 6 de 10 respuestas en un examen de tipo falso-verdadero.

7. Un agente de una compañía de seguros vende pólizas a 5 personas, todas de edad idéntica y con buena salud. De acuerdo con las tablas de actuarios la probabilidad de que una persona de esta edad específica esté viva en 30 años es de 2/3. Hallar la probabilidad de que en 30 años:

   a) Las 5 personas estén vivas.
   b) Al menos 3 estén vivas.
   c) Solamente 2 estén vivas
   d) Por lo menos 1 esté viva.

8. Sea X la variable aleatoria número de plantas con mutación en un total de 1000 plantas irradiadas, y p= 0.0001 la probabilidad de que una planta irradiada presente mutación. Se le pide calcular, usando la distribución de Poisson:

   a) La probabilidad de que no aparezca alguna planta con mutación.
   b) La probabilidad de que aparezca por lo menos una planta con mutación.
   c) El número medio (E[X]) de plantas con mutación.

9. Se sabe que dos pacientes de cada 1000 reaccionan a la penicilina. Si el día de hoy se someten 2000 pacientes a la prueba, calcule las siguientes probabilidades:

   a) Que 3 tengan reacción alérgica.
   b) Que más de 2 individuos tengan reacción.
   c) Calcule la E [X] y la Var [X].

10. Según la National Office of Vital Statistics of the US Department of Health, Education and Welfare, el promedio de ahogados por año es de 3.0 por cada 100,000 habitantes. Hallar la probabilidad de que en una ciudad cuya población es de 200,000 ocurran:

   a) 0 ahogados por año.
   b) 2 ahogados por año.
   c) 8 ahogados por año.
   d) Entre 4 y 8 ahogados por año.
   e) Menos de 3 ahogados por año.

11. Sí se sabe que en una cierta región ocurre en promedio una crecida de 550 m^3/seg a cada 20 años, calcule:

   a) La probabilidad de que ocurran dos o más crecidas en un año.
   b) La probabilidad de que no ocurra alguna creciente en un año.
   c) La probabilidad de que ocurran dos o más crecidas en 10 años.
   d) La probabilidad de que no ocurra alguna crecida en 10 años.
   e) La varianza y el valor esperado del número de crecidas en 20 años.

12. Sea X la variable aleatoria número de higos inapropiados para el consumo por caja con un cierto número de higos. Suponiendo que la probabilidad *p* de que un higo sea inapropiado para el consumo sea igual a 0.1, calcule:

   a) La probabilidad de que una caja con 8 higos, escogida al azar:
      - No contenga higos inapropiados para el consumo,
      - Contenga a lo sumo 1 higo inapropiado para el consumo.
   b) La probabilidad de que una caja con 9 higos, escogida al azar:
      - No contenga higos inapropiados para el consumo,
      - Contenga un máximo de 1 higo inapropiado por caja.
   c) El número medio esperado de higos inapropiados para el consumo por caja con 8 higos y Var (X).

13. Un director de producción sabe que el 5% de las piezas producidas en cierto proceso de fabricación tiene algún defecto. Se examinan seis de estas piezas, cuyas características se asumen independientes.

a) ¿Cuál es la probabilidad de que ninguna de estas piezas tenga un defecto?
b) ¿Cuál es la probabilidad de que una de estas piezas tenga un defecto?
c) ¿Cuál es la probabilidad de que al menos dos de estas piezas tengan un defecto?

14. Si el 3% de las lámparas eléctricas producidas por una compañía son defectuosas, encuentre la probabilidad de que en una muestra de 100 lámparas eléctricas, hallan exactamente:

a) 0
b) 1
c) 2 } lámparas defectuosas
d) 3
e) 5
f) entre 1 y 3 lámparas defectuosas
g) más de 5 lámparas defectuosas

15. Un vendedor de seguros sabe que la oportunidad de vender una póliza es mayor mientras más contactos realice con clientes potenciales. Si la probabilidad de que una persona compre una póliza de seguro después de la visita, es constante e igual a 0.25, y si el conjunto de visitas constituye un conjunto independiente de ensayos, ¿cuántos compradores potenciales debe visitar el vendedor para que la probabilidad de vender por lo menos una póliza sea de 0.80?

16. De cada 2,000 personas a las que se suministra cierto medicamento 6 resultan alérgicas al mismo, por término medio. Si en un determinado día se ha administrado el medicamento a 400 personas, ¿cuál es la probabilidad de que haya al menos una alérgica?

17. La probabilidad de que un golfista haga hoyo en un cierto tipo de lanzamiento es 0.2. Si lo intenta 5 veces, calcular la probabilidad de que:

a) no acierte alguna vez;
b) acierte por lo menos dos veces.
c) Supongamos que lanzara 10,000 veces y su capacidad de acierto se mantuviera (ni aumentara por la práctica ni disminuyera por el cansancio). ¿Qué probabilidad hay de que acierte más de 2.080 veces?

18. Un lote de semillas de *Eucalyptus saligna* con una proporción de 5% de semillas híbridas (*E. saligna* × *E. cloeziana*) fue utilizado para la plantación de área. Sí diez árboles de esta área fueran seleccionadas al azar, cuál es la probabilidad de que:

a) ninguna de ellos sea híbrido;
b) por lo menos uno de ellos sea híbrido;
c) todos sean híbridos.

19. En un bosque de *Eucalyptus grandis* la tasa de ocurrencia de cáncer es de 2.5%. En un inventario forestal fueron seleccionados al azar 30 árboles, ¿cuál es la probabilidad de que por lo menos un árbol tenga cáncer?

20. Una empresa agroexportadora ubicada en el valle de la Fragua, Zacapa, asegura que el 90% de los melones embarcados están maduros y listos para comer. Calcule la probabilidad de que entre 18 melones embarcados, por lo menos 16 estén maduros y listos para comer. Calcule la esperanza matemática e interprétela.

21. En promedio, cinco pájaros chocan contra el monumento a Washington y mueren por este motivo cada semana. Bill Garcy, un oficial del Servicio de Parques Nacionales de Estados Unidos, ha solicitado que el Congreso estadounidense asigne fondos para adquirir equipo que aleje a los pájaros del monumento. Un Subcomité del Congreso le ha respondido que no pueden asignarle fondos para tal fin a menos que la probabilidad de que mueran más de tres pájaros cada semana sea mayor a 0.7. ¿Deben destinarse los fondos para espantar pájaros?

22. La concertista de piano Donna Prima está muy molesta por el número de tosidos que se presentan en la audiencia justo antes que empiece a tocar. Durante su última gira, Donna estimó un promedio de ocho tosidos justo antes de empezar su concierto. La señora Prima le ha advertido a su director que si escucha más de cinco tosidos en el concierto de esa noche, se rehusará a tocar. ¿Cuál será la probabilidad de que la artista toque esa noche?

23. Un director regional tiene la responsabilidad del desarrollo de una empresa, y le preocupa la cantidad de quiebras de empresas pequeñas. Si la cantidad promedio de quiebras de empresas pequeñas por mes es de 10, ¿cuál es la probabilidad de que quiebren exactamente cuatro empresas pequeñas durante un mes? Suponga que la probabilidad de una quiebra es igual en dos meses cualesquiera, y que la ocurrencia o no ocurrencia de una quiebra en cualquier mes es independiente de las quiebras en los demás meses.

24. Un supervisor de seguridad en una empresa cree que el número esperado de accidentes laborales por mes es de 3.4
a) ¿Cuál es la probabilidad de que el próximo mes ocurran exactamente dos accidentes?
b) ¿Cuál es la probabilidad de que el próximo mes ocurran tres o más accidentes?
c) ¿Qué supuestos debe hacer usted para resolver estas preguntas mediante la distribución Poisson?

25. Como una forma de hacer control de calidad en una empresa comercializadora de puertas de madera, el dueño exige que antes de salir de la fábrica cada puerta sea revisada en busca de imperfecciones en la superficie de madera. El encargado de control de calidad encontró que el número medio de imperfecciones por puerta es 0.5. El dueño decidió que todas las puertas con dos o más imperfecciones sean rechazadas y sean devueltas para su reparación.

a) ¿Cuál es la probabilidad de que una puerta falle la inspección y sea devuelta para su reparación?
b) ¿Cuál es la probabilidad de que una puerta pase la inspección?

26. El número medio de pacientes admitidos por día en la sala de emergencias de un hospital pequeño es 2.5. Si solo hay cuatro camas disponibles en dicha sala ¿cuál es la probabilidad de que un día cualquiera el hospital no tenga camas suficientes para acomodar a los pacientes que lleguen?

27. Las últimas estadísticas de salud, afirman que en la zona del oriente antioqueño (en Colombia) se presenta una alta incidencia de cáncer de estómago (120 casos por cada 100,000 habitantes). Suponga que se realizan exámenes a 1000 habitantes del municipio de Guarne y se asume que para éstos la tasa de incidencia es la misma que para toda la región del oriente antioqueño.

a) ¿Cuál es la probabilidad de que ninguna de las personas examinadas tenga cáncer?
b) ¿Cuál es la probabilidad de que al menos 3 personas tengan cáncer?
c) ¿Cuál es la probabilidad de que al menos 8 personas tengan cáncer?

28. Un productor de naranjas tiene dos alternativas para la venta de su producto, que es empacado en cajas de 10 docenas.

a) Un comprador A que paga a US$10 la caja y no examina el producto;

b) Un comprador B que para cada caja recibida, retira 6 naranjas y las examina: si todas están perfectas él paga US$12 la caja; si entre las 6 encuentra 1 deteriorada, él paga US$10 la caja y, si entre las 6 encuentra de 2 a 4 deterioradas, él paga apenas US$6 por la caja; si encuentra más de 4 deterioradas él descarta la caja.

Si el porcentaje real de naranjas deterioradas es de 8%, ¿cuál es la mejor alternativa para la venta del producto? Y si el porcentaje de naranjas deterioradas fuera de 15%, ¿qué pasaría?

29. En una carretera de poco movimiento, pasa en promedio, 1 carro cada 20 minutos. Calcule la probabilidad de que en 30 minutos pasen:

   d) Ningún carro,
   e) Por lo menos 2 carros.

Dos apostadores se colocan en un punto estratégico de la carretera y hacen la siguiente apuesta, referente al número de carros que pasan en 1 hora de observación:
1 carro, A recibe de B US$ 20,
2 carros, A recibe de B US$30,
3 carros, A recibe de B US$40,
4 carros, A recibe de B US$ 50.
Si no pasa algún carro, A paga a B US$100. Si pasan más de 4 carros, la apuesta no tiene validez.

¿Cuál es la ganancia media del apostador A?

Sugerencia: Admita una distribución de Poisson.

30. Se estima que el 70 % de una población de consumidores prefiere una marca en particular de pasta de dientes A ¿Cuál es la probabilidad que al entrevistar a un grupo de consumidores.

a) sea necesario entrevistar exactamente 4 personas para encontrar el primer consumidor que prefiere la marca A? R/ 0.0189

b) Se tenga que entrevistar a lo más 6 personas para encontrar el primer consumidor que prefiere la marca A? R/ 0.00243

31. La probabilidad de que una muestra de aire contenga una molécula rara es de 0.01 si se supone que las muestras son independientes con respecto a la presencia de la molécula rara.

¿Cuál es la probabilidad de que sea necesario analizar exactamente 125 muestras antes de detectar una molécula rara? R/ 0.0029

32. Sea una máquina despachadora de refrescos que arroja un poco más de 20 ml por vaso derramándose el líquido en un 5% de los vasos despachados. Podemos definir la variable aleatoria X: “cantidad de vasos despachados hasta obtener el primero que se derramará” Considere que la forma de despachar el líquido por la máquina es independiente de vaso en vaso.

a) calcule la probabilidad de que el primer vaso que se derrame se encuentre después del 15vo. vaso despachado. R/0.4632

b) ¿Qué vaso despachado se espera sea el primero en el que se derrame el líquido? R/ 20

33. Un inspector de la SECOFI ha encontrado que en 6 de 10 tiendas que visita se presentan irregularidades. Si el inspector visita una serie de tiendas al azar ¿Cuál es la probabilidad de que:

a) la primera tienda con irregularidades fuera encontrada después de revisar la cuarta? R/ 0.0256

b) ¿cuántas tiendas se espera que tenga que visitar para encontrar la primera con irregularidades? R/1.666

34. En el salón hay 12 mujeres y 17 hombres si extraemos alumno tras alumno hasta completar 7 mujeres ¿Cuál es la probabilidad de que hayan salido 2 hombres? R/ 0.0199

35. Si la probabilidad de que un niño expuesto a una enfermedad contagiosa la contraiga es 0.40, ¿Cuál es la probabilidad de que el décimo niño expuesto a la enfermedad sea el tercero en contraerla? R/ 0.0645

36. Para evitar que lo descubran en la aduana, un viajero ha colocado 6 tabletas de narcótico en una botella que contiene 9 píldoras de vitamina que son similares en apariencia. Si el oficial de la aduana selecciona 3 tabletas aleatoriamente para analizarlas, a) ¿Cuál es la probabilidad de que el viajero sea arrestado por posesión de narcóticos?, b) ¿Cuál es la probabilidad de que no sea arrestado por posesión de narcóticos?

37. a)¿Cuál es la probabilidad de que una mesera se rehúse a servir bebidas alcohólicas únicamente a dos menores de edad si verifica aleatoriamente solo 5 identificaciones de entre 9 estudiantes, de los cuales 4 no tienen la edad suficiente?, b) ¿Cuál es la probabilidad de que como máximo 2 de las identificaciones pertenezcan a menores de edad?

38. Una compañía manufacturera utiliza un esquema para la aceptación de los artículos producidos antes de ser embarcados. El plan es de dos etapas. Se preparan cajas de 25 para embarque y se selecciona una muestra de 3 para verificar si tienen algún artículo defectuoso. Si se encuentra uno, la caja entera se regresa para verificarla al 100%. Si no se encuentra ningún artículo defectuoso, la caja se embarca.

a)¿Cuál es la probabilidad de que se embarque una caja que tiene tres artículos defectuosos?, b)¿Cuál es la probabilidad de que una caja que contiene solo un artículo defectuoso se regresa para verificación?

3.13 DISTRIBUCIONES DE PROBABILIDAD CONTINUAS

3.13.1 FUNCIÓN DENSIDAD DE PROBABILIDAD

Se dice que f(x) es una función continua de probabilidad o función densidad de probabilidad para una variable aleatoria continua X, si satisface dos condiciones:

- $f(x) \geq 0$, para todo $x \in (-\infty , \infty)$;
- El área definida por f(x) es igual a 1.

Con el apoyo de cálculo diferencial e integral, se puede verificar la segunda condición a través de:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

De la misma manera, para calcular probabilidades, se tiene que para $a \leq b$:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\, dx \qquad \text{Área bajo la función definida por el intervalo } [\,a, b\,]$$

Por la forma como se atribuyen las probabilidades en el caso continuo, se tiene que un área igual a 0 bajo cualquier valor individual, esto es, P(X=k) = 0 para cualquier k. Por tanto, en el caso de variables aleatorias continuas, la probabilidad de que X sea igual a cualquier valor determinado es igual a 0, y consecuentemente, las probabilidades calculadas sobre los intervalos [a, b], [a, b), (a, b] y (a, b) son las mismas, para cualquier valor de a y b.

Ejemplo 63

Un grupo de arqueólogos estudiaron una cierta región y establecieron un modelo teórico para la variable C, largo de los fósiles de la región (en centímetros). Suponga que C es una variable aleatoria continua con la siguiente función densidad de probabilidad:

$$f(c) = \begin{cases} \frac{1}{40}\left(\frac{c}{10}+1\right) & \text{si} \quad 0 \leq c \leq 20 \\ 0 & \text{caso contrario} \end{cases}$$

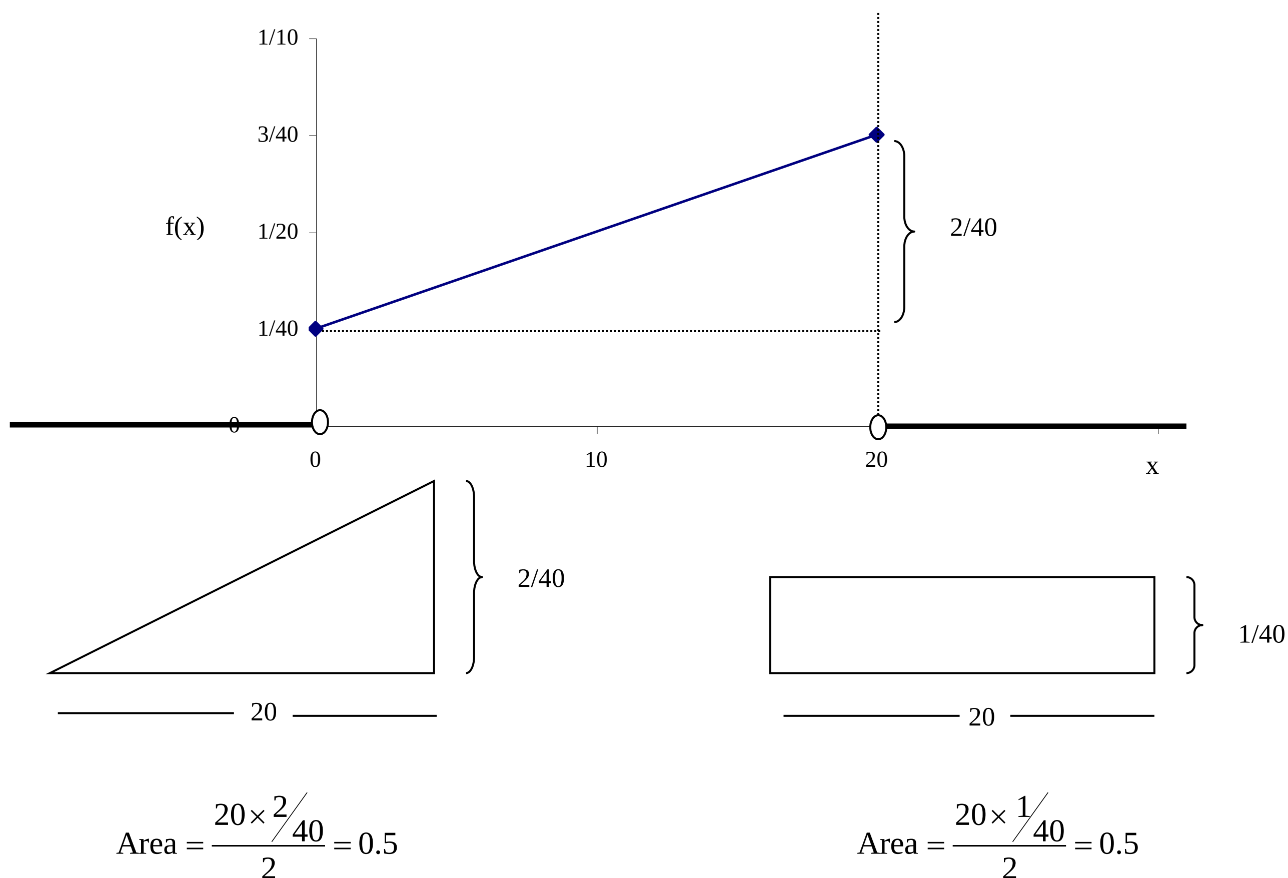
Verifique que f(c) es realmente una función densidad de probabilidad.

$$\frac{1}{40}\left(\frac{c}{10}+1\right) = \left(\frac{c}{400}+\frac{1}{40}\right) \Rightarrow \qquad \int_{0}^{20}\left(\frac{c}{400}+\frac{1}{40}\right)dc = 1 \qquad \Rightarrow$$

$$\int_0^{20} \left(\frac{c}{400}\right) dc + \int_0^{20} \left(\frac{1}{40}\right) dc \;=\; 1$$

$$\Rightarrow \quad \frac{1}{400}\frac{c^2}{2}\Bigg|_0^{20} + \frac{1}{40}c\Bigg|_0^{20} = 1 \quad \Rightarrow \quad \left(\frac{20^2}{800} - 0\right) + \left(\frac{20}{40} - 0\right) = 1 \quad \Rightarrow \quad 0.5 + 0.5 \;=\; 1$$

En forma gráfica:



$$\text{Area} = \frac{20 \times {}^{2}\!/_{40}}{2} = 0.5 \qquad\qquad \text{Area} = \frac{20 \times {}^{1}\!/_{40}}{2} = 0.5$$

3.13.2 ESPERANZA MATEMÁTICA

El valor esperado o media de la variable aleatoria continua X, con función densidad dada por f (x) es dada por la siguiente expresión:

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

3.13.3 VARIANZA

Para una variable aleatoria X con densidad f(x), la varianza es dada por la siguiente expresión:

$$V(X) = \sigma^2_X = \int_{-\infty}^{\infty} (x - \mu_X)^2 \, f(x)\, dx$$

Como en el caso discreto, la varianza es la medida de dispersión más utilizada en la práctica. Aquí también se puede utilizar la expresión alternativa:

$$V(X) = \sigma^2_X = E(X^2) - \mu_X{}^2$$

Con $E(X^2)$ calculada de la siguiente forma:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \, f(x)\, dx$$

Ejemplo 64

Para la función densidad de probabilidad de la variable aleatoria continua C, largo de fósiles (en cm):

$$f(c) = \begin{cases} \frac{1}{40}\left(\frac{c}{10} + 1\right) & \text{si} \quad 0 \le c \le 20 \\ \\ 0 & \text{caso contrario} \end{cases}$$

Calcule E(C):

$$E(C) = \int_0^{20} c \, \frac{1}{40}\left(\frac{c}{10} + 1\right) dc \;\; = \;\; \frac{1}{400}\frac{c^3}{3}\Bigg|_0^{20} + \frac{1}{40}\frac{c^2}{2}\Bigg|_0^{20} = \frac{20}{3} + 5 = \frac{35}{3} = 11.67 \text{ cm}$$

Calcule V(C):

$$E(C^2) = \int_0^{20} c^2 \, \frac{1}{40}\left(\frac{c}{10} + 1\right) dc \;\; = \;\; \frac{1}{400}\frac{c^4}{4}\Bigg|_0^{20} + \frac{1}{40}\frac{c^3}{3}\Bigg|_0^{20} = 100 + \frac{200}{3} = \frac{500}{3}$$

Entonces: $V(C) = \sigma^2_C = E(C^2) - \mu_C{}^2 = \frac{500}{3} - \left(\frac{35}{3}\right)^2 = \frac{275}{9} = 30.56 \text{ cm}^2$

3.13.4 FUNCIÓN DISTRIBUCIÓN ACUMULADA

La función de distribución acumulada de probabilidad, denotada por F (x) para una variable aleatoria continua se define por:

$$F(x) = P(X \le x) = P(-\infty \le X \le x) = \int_{-\infty}^{x} f(u)\, du$$

Ejemplo 65

a) Encuentre el valor de la constante c para que la función:

$$f(x) = \begin{cases} c\,x^2 & \text{si} \quad 0 < c < 3 \\ 0 & \text{caso contrario} \end{cases}$$

b) Calcule P(1<X<2).

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \Rightarrow \int_0^3 cx^2\,dx = 1 \Rightarrow c\left.\frac{x^3}{3}\right|_0^3 = 1 \Rightarrow 9c = 1 \Rightarrow c = 1/9$$

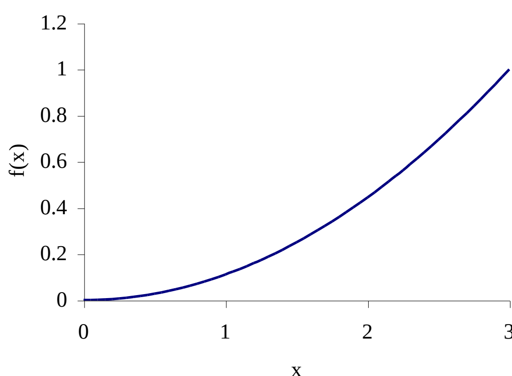
$$P(1 < x < 2) = \int_1^2 1/9\, x^2\,dx \Rightarrow \left.\frac{x^3}{27}\right|_1^2 = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}$$

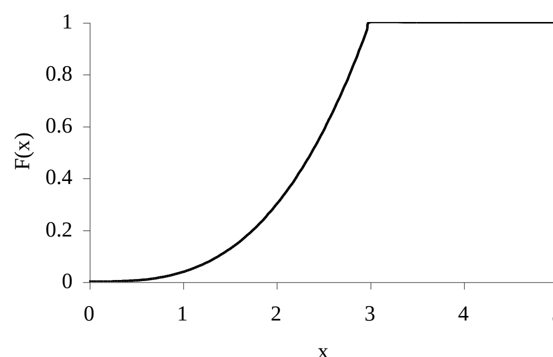
c) Encuentre la función de distribución acumulada

$$F(x) = \int_0^x 1/9\, u^2\,du \Rightarrow \left.\frac{u^3}{27}\right|_0^x = \frac{x^3}{27}$$

Por lo tanto, la función de distribución acumulada es:

$$F(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ x^3/27 & 0 \le x < 3 \\ 1 & x \ge 3 \end{cases}$$





LISTA DE EJERCICIOS 6

1. Suponga que el peso de recién nacidos (en kg) pueda ser considerado una variable aleatoria con la siguiente función densidad de probabilidad:

$$f(x) = \begin{cases} \frac{1}{10}x + \frac{1}{10}, & \text{si} \quad 0 \le x \le 2; \\ -\frac{3}{40}x + \frac{9}{20} & \quad 2 < x \le 6; \\ 0 & \text{caso contrario} \end{cases}$$

¿Cuál es la probabilidad de que, al escoger al azar un niño, tenga peso:

a) Inferior a 3 kg.
b) Entre 1 y 4 kg
c) Por lo menos 3 kg

2. El incremento anual en el área afectada por una cierta plaga, en una región productora de frutas puede ser modelado por una variable aleatoria continua, medida en hectáreas, con función densidad de probabilidad:

$$f(x) = \begin{cases} \frac{2}{3}x, & \text{si} \ 0 < x < 1; \\ 1 - \frac{x}{3}, & \text{si} \ 1 \le x < 3; \\ 0 & \text{caso contrario} \end{cases}$$

a) Construya el gráfico de esa función densidad de probabilidad.
b) ¿Cuál sería la probabilidad de que la plaga afecte entre 2 y 3 hectáreas por año?
c) Determine el incremento promedio E(X) anual en el área afectada por la plaga.

3. Una variable aleatoria continua X tiene función densidad de probabilidad:

$$f(x) = \begin{cases} cx^2, & \text{si} \ 1 \le x \le 2; \\ cx, & \text{si} \ 2 < x < 3; \\ 0 & \text{caso contrario} \end{cases}$$

a) Encuentre el valor de la constante c
b) Calcule P(X>2)
c) Calcule P(1/2<X<3/2)
d) Encuentre la función de distribución acumulada de la variable aleatoria X.

4. El consumo de combustible de un tipo de automóvil es una variable aleatoria medida en km/litro. Admita que la función densidad de probabilidad de esa variable es expresada por la siguiente función:

$$f(x) = \begin{cases} x - 10, & \text{si } 10 \le x \le 11; \\ 12 - x, & \text{si } 11 < x \le 12; \\ 0 & \text{caso contrario} \end{cases}$$

a) Construya el gráfico de la función densidad de probabilidad
b) Encuentre E(X) y V(X)
c) Siendo R$0.70 el precio del litro de combustible, ¿cuál será la media del gasto en un viaje de 100 km con ese automóvil?

5. Suponga que una variable aleatoria continua X tiene función densidad de probabilidad dada por:

$$f(x) = \begin{cases} \frac{1}{6}, & \text{si } 0 \le x \le 1; \\ \frac{1}{2}x - \frac{1}{3}, & \text{si } 1 < x \le 2; \\ -\frac{1}{2}x + \frac{5}{3}, & \text{si } 2 < x \le 3; \\ 0 & \text{caso contrario.} \end{cases}$$

a) Determine E(X)
b) Calcule V(X)

3.13.5 DISTRIBUCIÓN NORMAL

Uno de los más importantes ejemplos de una distribución de probabilidad continua es la distribución normal. La distribución Normal es también conocida como "distribución Gaussiana" como homenaje a Karl F. Gauss (1777-1855), brillante matemático e físico alemán, que la desarrollo a inicios del siglo XIX. Sin embargo, Abraham de Moivre (1667-1754) fue el primero en anunciar la ecuación de la distribución en el año 1733 y Pierre-Simon Marquis de Laplace (1749-1827), famoso matemático y físico francés, la redescubrió en la misma época que Gauss.

Para evitar "un problema internacional de originalidad" el famoso estadístico inglés Karl Pearson pasó a llamarla: distribución "Normal" en el año 1920.

Algunos ejemplos de variables aleatorias continuas que generalmente siguen una distribución normal son:

(a) Peso de: animales, personas, plantas, sacos de café, etc.,
(b) Altura de: árboles, personas, plantas, etc.
(c) Diámetros de: árboles, tornillos, etc.
(d) Productividad de: caña de azúcar, soya, brócoli, hule, etc.
(e) Volumen de: madera, líquidos, etc.

Las principales razones que hacen de la distribución Normal el modelo más importante en la Bioestadística son:

1. Muchas variables biométricas tienden a tener distribución Normal.

2. La distribución de las medias muestrales de una variable cualquiera tienden a tener distribución Normal, aunque la variable en sí, no tenga distribución Normal.

3. Muchas pruebas y modelos estadísticos tienen como suposición la "normalidad de los datos", esto es, que los dados poseen distribución Normal.

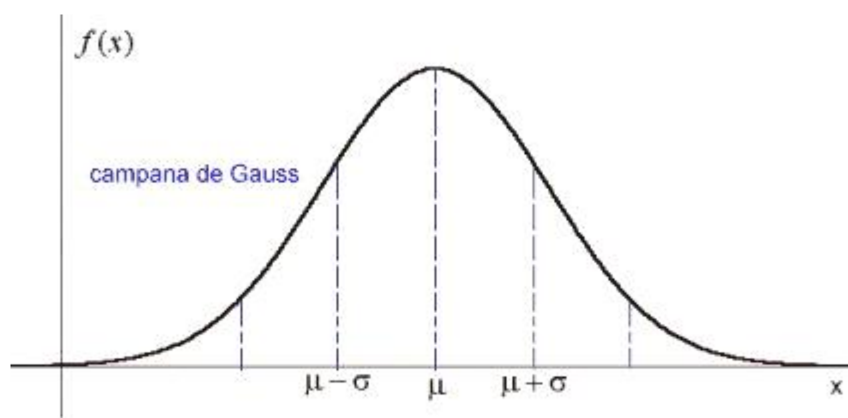
La variable aleatoria X tiene una distribución normal con parámetros μ ($-\infty < \mu < \infty$) y σ^2 ($\sigma^2 > 0$), sí su función densidad de probabilidad es dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)}{2\sigma^2}} \quad , \quad -\infty < x < \infty$$

siendo que:

e = base de los logaritmos neperianos o naturales, aproximadamente igual a 2.71828
π = 3.1415......
μ = medida poblacional (parámetro de localización)
σ^2 = varianza poblacional (parámetro de escala)

Si $X \sim N(\mu, \sigma^2)$ entonces el gráfico de la función densidad de probabilidad de X tiene la forma de una campana, como se muestra e continuación:



Algunas características de la función densidad de probabilidad:

i) Hay toda una familia de distribuciones normales de probabilidad. Cada distribución normal específica se distingue por su media μ y su desviación estándar σ.

ii) El punto más alto de la curva normal es la media que también es la mediana y la moda de la distribución.

iii) La distribución normal de probabilidad es simétrica

iv) Los dos extremos de la distribución normal de probabilidad se extienden indefinidamente y teóricamente nunca tocan el eje horizontal.

v) La desviación estándar determina el ancho de la curva. A valores mayores de la desviación estándar se tienen curvas más anchas y bajas, que muestran una mayor dispersión de los datos.

vi) Sin importar cuáles sean los valores de μ y σ, el área total bajo la curva de la distribución normal de probabilidad es 1.

vii) Las probabilidades de la variable aleatoria normal se determinan con las áreas bajo la curva.

viii) Todas las curvas de densidad de probabilidad normal satisfacen las siguientes propiedades que es frecuentemente referidas como *Regla Empírica.*

68% de las observaciones están comprendidas entre **1 desviación estándar** de la **media**, esto es, entre $\mu - \sigma$ y $\mu + \sigma$.

95% de las observaciones están comprendidas entre **2 desviación estándar** de la **media**, esto es, entre $\mu - 2\sigma$ y $\mu + 2\sigma$.

99% de las observaciones están comprendidas entre **3 desviación estándar** de la **media**, esto es, entre $\mu - 3\sigma$ y $\mu + 3\sigma$.

Ejemplo 66

Sabiendo que una variable aleatoria X, diámetro (en mm) de un tomate tiene una distribución normal N (60,49), calcule:

a) $P(X<50)$
b) $P(40 < X < 55)$

PROBLEMA: No existe una función primitiva para la función densidad de probabilidad normal

- Una solución: Utilizar métodos de cálculo aproximado para de integrales definidas.

- Otra solución: consultar una tabla de probabilidades

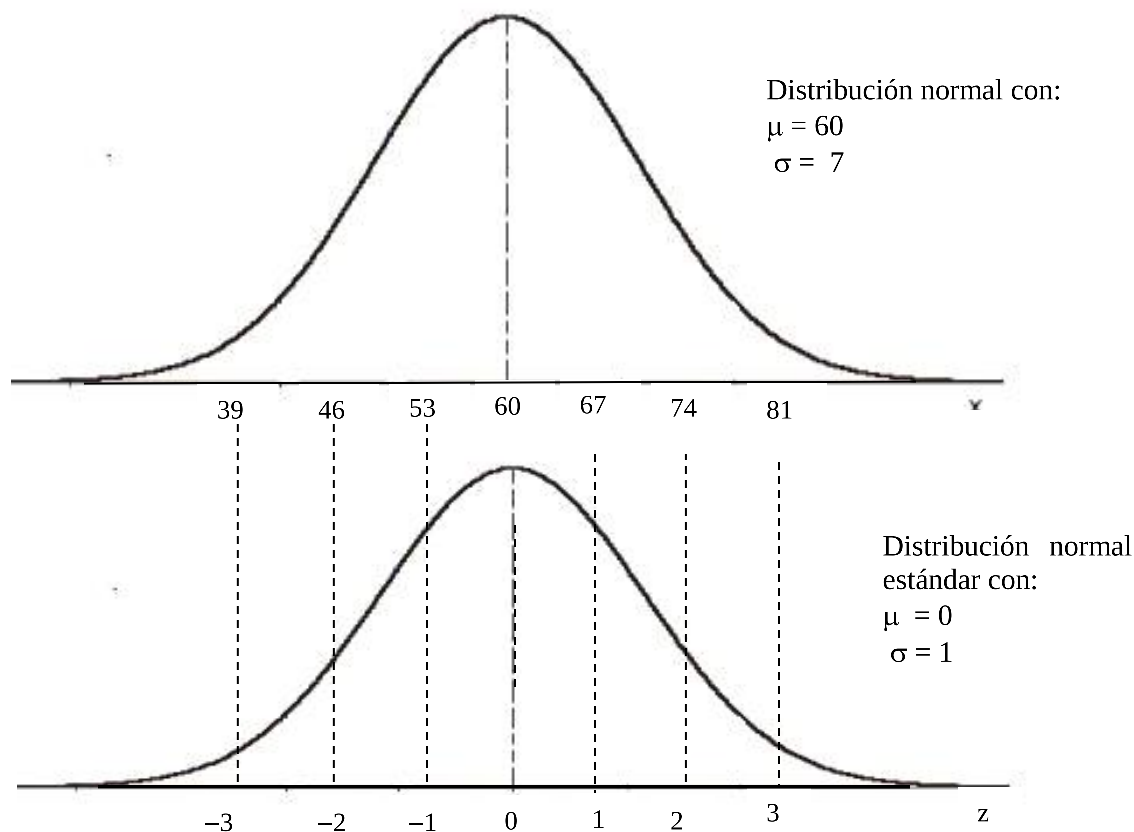
NUEVO PROBLEMA: Sólo existe una tabla, con las probabilidades para la distribución normal con media μ=0 y varianza σ^2=1, llamada: DISTRIBUCIÓN NORMAL ESTANDAR o REDUCIDA.

- Solución: Transformar la variable X en $Z = \frac{x-\mu}{\sigma}$, en la que:

x = valor de la variable aleatoria que nos preocupa.
μ = media de la distribución de la variable aleatoria.
σ = desviación estándar de la variable aleatoria.
Z = número de desviaciones estándar que hay de X a la media de la distribución.

Y luego utilizar la tabla de probabilidades de la distribución normal estándar. ¿Por qué utilizar Z?

Las variables aleatorias normalmente distribuidas tienen muchas unidades diferentes de medición: dólares, pulgadas, partes por millón, kilogramos, segundos, etc. Como se utilizará una sola tabla, se hablará en términos de unidades estándar (que en realidad significa desviaciones estándar) y se denota a éstas con el símbolo Z. Lo anterior se puede demostrar en forma gráfica. En la siguiente figura se puede observar que el uso de Z es solamente un cambio en la escala de medición del eje horizontal.

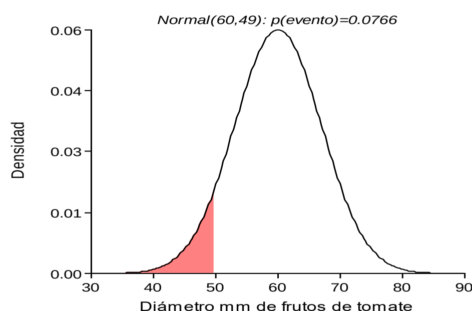


Solución al ejercicio planteado:

x = variable aleatoria, diámetro (en mm) de tomate.
μ = 60 mm.
σ = 7 mm.

Calcule:

a) P (X<50)



$$z = \frac{50-60}{7} = -1.43$$

$$P(z) = 0.4236 \ \ \text{(corresponde al area A)}$$

Consulte la tabla como se muestra a continuación.

Respuesta:

$$P(X < 50) = 0.5 - 0.4236 = 0.0764$$

DISTRIBUCIÓN NORMAL

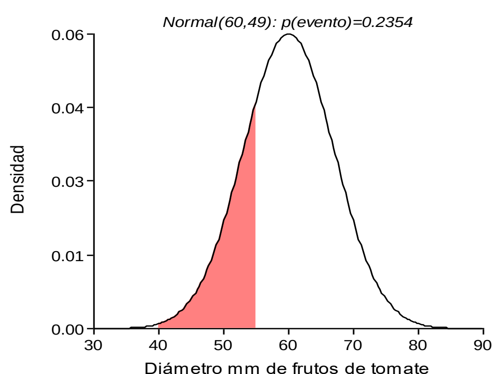
0 z x

Área bajo la curva Normal de 0 a Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.2737	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408

- Como la distribución normal es simétrica, el valor de 0 a 1.43 es equivalente a 0 a -1.43

b) P (40 < X < 55)



$$z_1 = \frac{40-60}{7} = -2.86$$

$$P(z_1) = 0.4979$$

$$z_2 = \frac{55-60}{7} = -0.71$$

$$P(z_2) = 0.2611$$

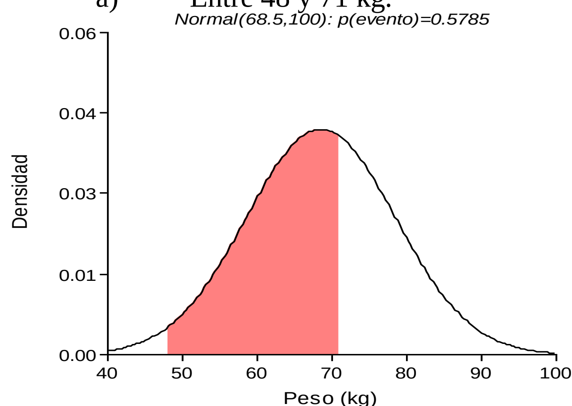
$$P(40 < X < 55) = 0.4979 - 0.2611 = 0.2368$$

c) P (X > 60) = 0.5

Ejemplo 67

El peso medio de 500 estudiantes varones de una universidad es de 68.5 kg y la desviación estándar es de 10 kg. Suponiendo que los pesos están distribuidos normalmente, hallar el número de estudiantes que pesan:

a) Entre 48 y 71 kg.



$$z_1 = \frac{48-68.5}{10} = -2.05$$

$$P(z_1) = 0.4798$$

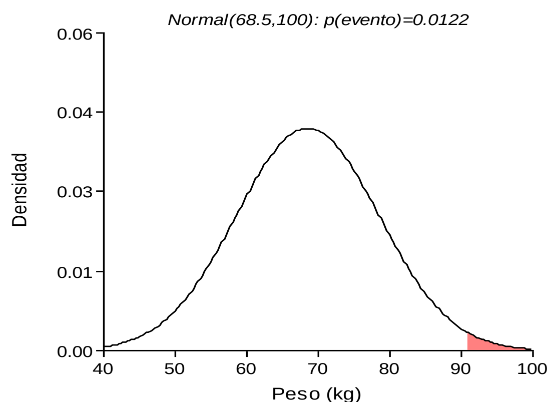
$$z_2 = \frac{71-68.5}{10} = 0.25$$

$$P(z_2) = 0.0987$$

$$P(40 < X < 55) = 0.4798 + 0.0987 = 0.5785$$

El número de estudiantes que pesan entre 48 y 71 kg es 500 (0.58) = 290.

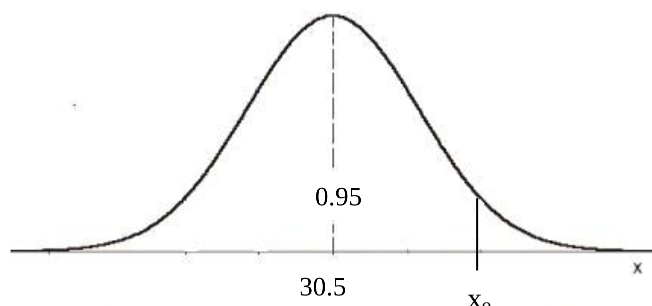
b) Más de 91 kg.



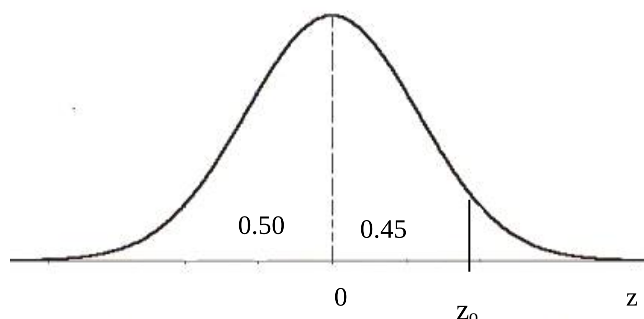
Ejemplo 68

Ciertos estudios muestran que el rendimiento de gasolina para los automóviles compactos, vendidos en los EE.UU. tienen una distribución normal, con un rendimiento medio de 30.5 millas por galón (mpg) y una desviación estándar de 4.5 mpg. Si un fabricante desea diseñar un auto compacto más económico que el 95% de los autos compactos vendidos en los EE.UU. ¿Cuál debe ser el rendimiento mínimo del auto nuevo?

Solución: Sea $X \sim N(30.5\ ;\ 4.5)$, se desea encontrar el valor de x_o tal que: $P(X< x_o) =0.95$



1º Se encuentra el valor de z_o que corresponde a x_o, es decir, el valor de z_o tal que el área a su izquierda sea igual a 0.95. Puesto que el área a la izquierda de z_o=0 es 0.5, z_o será el valor de z tabular que corresponde a un área igual a 0.45. Este valor es z_o =1.645



2º Se debe encontrar el valor de x_o correspondiente a z_o=1.645:

$$1.645 = \frac{x_o - 30.5}{4.5} \qquad \Rightarrow \quad x_o = (4.5)(1.645) + 30.5 = 37.9$$

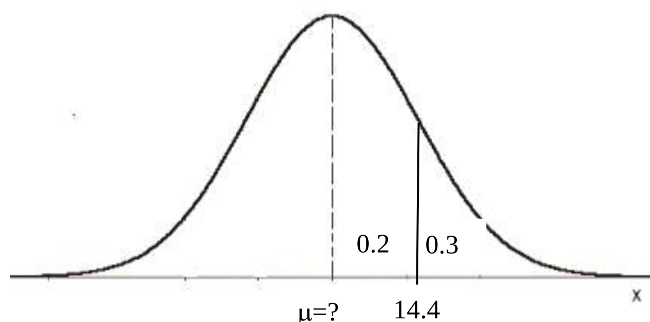
3º El nuevo auto compacto del fabricante debe desarrollar un tener un rendimiento de 37.9 mpg (por lo menos) para ser mejor que el 95% de los autos compactos que actualmente se venden en los Estados Unidos.

Ejemplo 69

Una variable X tiene una distribución normal con una media μ desconocida y una desviación estándar σ = 1.8. Si la probabilidad de que x sea mayor que 14.4 es 0.3, encuentre el valor de μ.

Solución:

1º Se debe encontrar el valor de z que corresponde a P(z)=0.2



Para un valor de P(z)=0.2, corresponde un valor de z de 0.5244.

2º Se sustituye el valor de z en la ecuación: $z = \dfrac{x-\mu}{\sigma}$ y se tiene que:

$$0.5244 = \frac{14.4-\mu}{1.8} \quad \Rightarrow \quad \mu = 14.4-(0.5244)(1.8) \quad \Rightarrow \mu = 13.456$$

3.13.6 APROXIMACIÓN NORMAL PARA LA DISTRIBUCIÓN DE PROBABILIDAD BINOMIAL

La distribución de probabilidad binomial se aproxima utilizando una curva normal (teorema de De Moivre-Laplace), con:

$$\mu = np$$
$$\sigma = \sqrt{npq}$$

Siendo:
n = número de ensayos o pruebas.
p = probabilidad de éxito en un solo ensayo
q = 1 – *p*, o sea, probabilidad de fracaso.

La aproximación será adecuada cuando:

a) La probabilidad *p* de ocurrir un éxito no está muy próxima de 0 ó de 1.

b) El número *n* de ensayos es grande ($n \geq 20$), de tal forma que $np \geq 5$ y $n(1-p) \geq 5$

c) El intervalo $\mu \pm 2\sigma$ está entre 0 y *n*.

Ejemplo 70

En 10 lanzamientos de una moneda honrada, hallar la probabilidad de obtener:

a) Entre 3 y 6 caras inclusive,
b) Exactamente 7 caras,
c) Más de 4 caras.

Utilizando la distribución binomial y la aproximación normal.

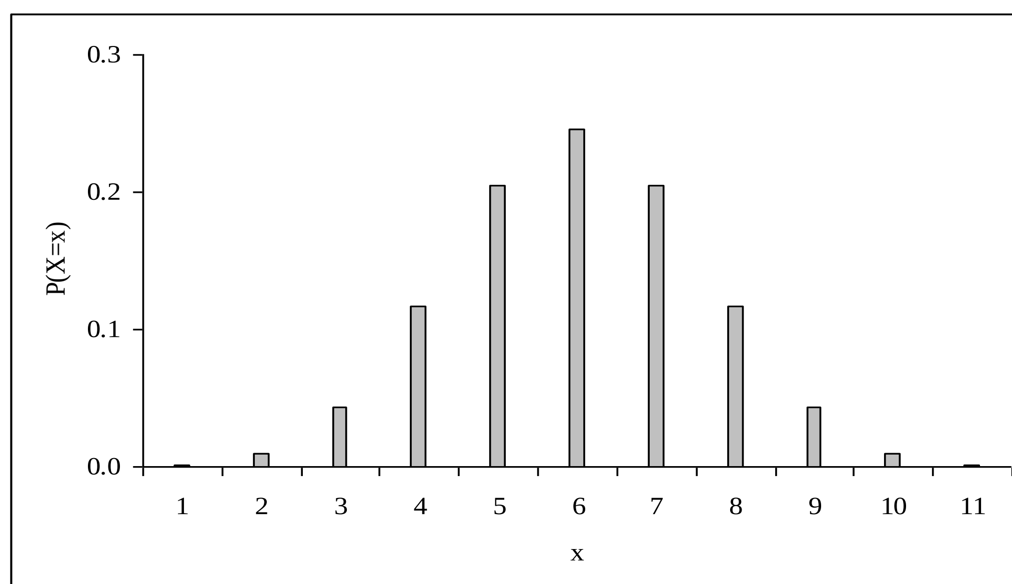
Solución:

(a) Sea X la variable aleatoria que da el número de caras en 10 lanzamientos. Entonces, utilizando la distribución binomial, se tiene:

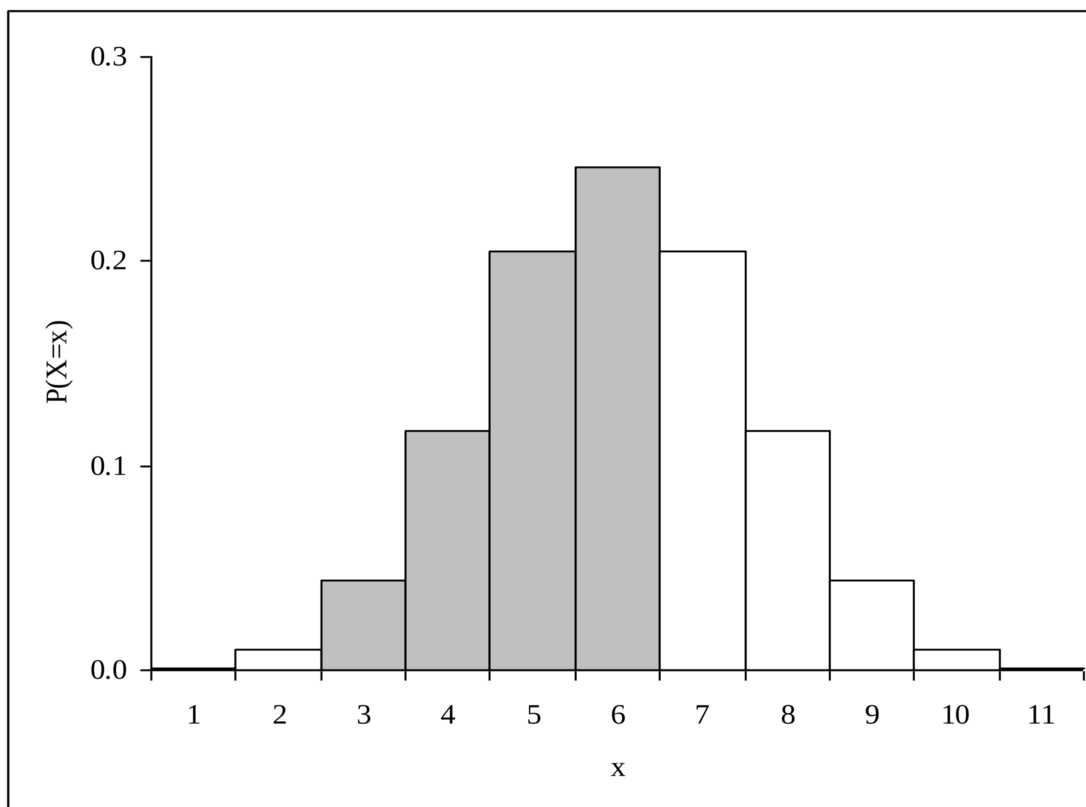
X	P(X=*x*)
0	0.001
1	0.010
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.010
10	0.001

$P(3 \leq X \leq 6) =$ **0.7734**

Ahora, para utilizar la distribución normal, se ilustrará construyendo la distribución de probabilidad para este experimento.



La figura siguiente muestra la distribución, como si fuera contínua:



La probabilidad pedida es la suma de las áreas de los rectángulos sombreados en la figura anterior, y puede aproximarse por el área bajo la correspondiente curva normal. Considerando los datos como continuos, se deduce que 3 a 6 caras pueden considerarse como: 2.5 a 6.5 caras. Ese mecanismo que consiste en alterar en 0.5 unidades el valor con que se desea calcular la probabilidad, se denomina: corrección de continuidad. Además, la media y la desviación estándar para la distribución binomial están dadas por:

$$\mu = np = 10\left(\tfrac{1}{2}\right) = 5$$

$$\sigma = \sqrt{npq} = \sqrt{10\left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right)} = 1.58$$

Entonces X ~ N (5, 1.58), y

- 2.5 en unidades tipificadas = $Z_1 = \frac{x_i - \mu}{\sigma} = \frac{2.5 - 5}{1.58} = -1.58$

  $P(Z_1) = P(-1.58 \le Z \le 0) = 0.4429$

- 6.5 en unidades tipificadas = $Z_2 = \frac{x_i - \mu}{\sigma} = \frac{6.5 - 5}{1.58} = 0.95$

  $P(Z_2) = P(0 \le Z \le 0.95) = 0.3289$

Entonces P (3 ≤ X ≤ 6) = 0.4429 + 0.3289 = **0.7718**

Nota: La precisión es mejor para valores superiores de *n*.

(b) P (X=7) = **0.117** (utilizando la distribución binomial).

Por la aproximación normal se tiene: P (6.5 ≤ X ≤ 7.5), en consecuencia:

- $Z_1 = \frac{x_i - \mu}{\sigma} = \frac{7.5-5}{1.58} = 1.58$
  P (Z_1) = 0.4429

- $Z_2 = \frac{x_i - \mu}{\sigma} = \frac{6.5-5}{1.58} = 0.95$
  P (Z_2) = 0.3289

Entonces P (6.5 ≤ X ≤ 7.5) = 0.4429 – 0.3289 = **0.114**

(c) P(X > 4)

c.1 Binomial: $\sum_{i=5}^{10} P(x_i) = 0.623$

c.2 Aproximación normal P (X > 4) ≈ P (X ≥ 4.5)

$Z = \frac{x_i - \mu}{\sigma} = \frac{4.5-5}{1.58} = -0.316$ P (Z) = 0.1255

P (X ≥ 4.5) = P (Z ≥ –0.316) = 0.5 +0.1255 = **0.626**

Corrección de continuidad o de Yates:

Cuando aproximamos una distribución binomial mediante una normal, estamos convirtiendo una variable X discreta (toma un número determinado de valores) en una continua **X'** (toma valores en un intervalo). Los valores de la probabilidad para valores fijos de la variable continua son cero (ya que sería el área de un punto), y necesitamos definir un intervalo.

Para evitar este problema en la aproximación de los valores fijos estos se corrigen (corrección de continuidad o de Yates) sustituyéndolos por un intervalo centrado en el punto y de valor unidad. En el siguiente esquema se muestran todas las situaciones posibles:

X⇒B(n,p) y **X'**⇒N(n·p, $\sqrt{n.p.q}$)

P(X=a)=P(a-0.5≤**X'**≤a+0.5)
P(X≤a)=P(**X'**≤a+0.5) (para que contenga al punto a)
P(X<a)=P(X'a≤a-0.5) (para que no contenga al punto a)
P(X>a)=P(X'≥a+0.5) (para que no contenga al punto a)
P(X≥a)= P(X'≥a-0.5) (para que contenga al punto a)
P(a≤X<b) =P(a-0.5≤ **X'**≤ b+0.5) (para que contenga al punto a y no a b)

Ejemplo 71

En una granja avícola han observado que el peso (en gramos) de los pollos de cuatro semanas sigue una distribución normal de tipo N (μ, σ) con μ = 1030 gramos, σ = 50 gramos (σ^2= 2500). La inspección sanitaria considera que los pollos cuyo peso es inferior a $\mu - 1.5\sigma$ son no aptos, y deben ser apartados para recibir un tratamiento especial. Esta mañana, en una inspección de sanidad rutinaria, hemos elegido 100 pollos de cuatro semanas de esa granja (elección con reemplazamiento; una vez pesado el pollo se devuelve al corral y podría volver a ser elegido posteriormente). ¿Cuál es la probabilidad de que de esos 100 pollos haya al menos 10 no aptos?

Solución del ejercicio:

1. Calcule la probabilidad de encontrar un pollo no apto.

   La inspección sanitaria considera que los pollos cuyo peso es inferior a $\mu - 1.5\sigma$ son no aptos, y deben ser apartados para recibir un tratamiento especial. Por lo tanto Z = − 1.5, o si desea calcular el valor de probabilidad usando μ= 1030 gramos, se tiene que X = 1030-1.5 (50) = 955 gramos, ya que $X = \mu - 1.5\sigma$.

   Por tanto:

   $P(Z < -1.5) = 0.06680730458$

   $P(X<955) = 0.06680730458$

2. ¿Cuál es la probabilidad de que de esos 100 pollos haya al menos 10 no aptos?

a) Usando la distribución binomial: n = 100, p = 0.06680730458, q = 0.933133, $x \geq 10$ (que es equivalente a X >9). **R/ 0.13122**

b) Utilizando la aproximación normal (teorema de De Moivre-Laplace), se tiene que:

E(X) = μ = n.p = (100)(0.06680730458)= 6.68073,
Var (X) = n.p.q = = (100)(0.06680730458) (1- 0.06680730458) = 6.234409.

Y lo que deseamos calcular es P(X ≥ 10), o sea, P(X >9). Aplicado la corrección de Yates, tenemos los valores corregidos así: P(X ≥ 10) = P(X ≥ 9.5) = P(X >9.5). Usando Infostat se tiene P(X ≥ 10) = **0.1294243959.**

La diferencia con lo obtenido en el inciso a es 0.002

Calculador de probabilidades y cuantiles

Seleccione la distribución
Uniforme (a,b)
Normal (media,varianza)
T Student (v)
Chi Cuadrado (v,lambda)
F no central (u,v,lambda)
Exponencial (lambda)
Gama (lambda,r)
Beta (a,b)
Weibull (a,b)
Doble Exponencial (a,b)
Logística (a,b)
Pareto (Theta,a)
Gumbel (a,b)
Rangos estudent.(k,v)
Poisson (lambda)
Binomial (n,p)
Geométrica (p)
Hipergeométrica (m,k,n)
Binomial Negativa (m,k)
Beta-Binomial(p,Rho,N)

6.68073 media
6.234409 varianza
Valor de x
9.5
Prob. (X<=x)
0.8705756041
Prob. (X>x)
0.1294243959
Prob. (X=x)
0
Calcular
Ayuda

Ejemplo 72

Un granja provee pollos en cajas de 20 animales, con peso medio de los pollos igual a 1.5 kg y desviación estándar de 0.3 kg. El peso de cada caja es constante e igual a 2 kg. En un cargamento de 100 cajas, cuál es la probabilidad de que el peso total sea inferior a 3300 kg

Respuesta:

μ = 1.5 kg y σ = 0.3 kg

Sea X la variable aleatoria peso de los pollos, X ~N (1.5, 0.3)

Si se consideran muestras de tamaño n= 20 pollos, se tiene que el peso promedio de los pollos sigue una distribución normal con $\bar{x} = 1.5\, kg \quad y \quad \frac{\sigma}{\sqrt{n}} = \frac{0.3}{\sqrt{20}} = 0.06708204$, es decir,

$\overline{x} \sim N(1.5\,, 0.0671)$.

El peso total de cada caja (con 20 pollos) sigue también una distribución normal, con parámetros:

$$n\mu = 20*1.5 = 30\ kg,\ y$$

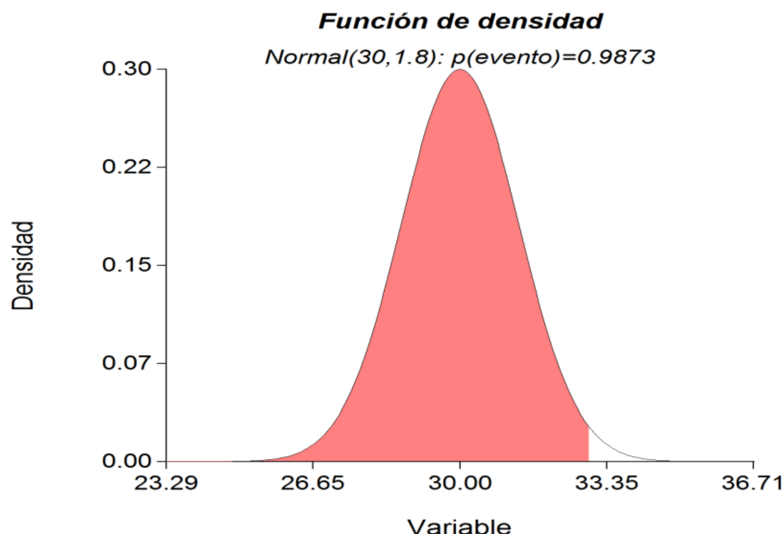
$$n.\frac{\sigma}{\sqrt{n}} = \sigma.\sqrt{n} = 0.3*\sqrt{20} = 1.3416\ kg.$$

$\hat{\tau} \sim N(30\,, 1.3416)$. La varianza es igual a : 1.3416^2=1.8

En un cargamento de 100 cajas, cuál es la probabilidad de que el peso total:

a) sea inferior a 3300 kg.,

a.1) se puede transformar todo a peso promedio por caja, por lo que 3300 kg/ 100 cajas = 33 kg/caja. El valor de probabilidad requerido es P(X<33 kg/caja). **R/ 0.9873 (98.73%)**



a.2) Transformando todo a peso:

$$N(n\mu) = 100*20*1.5 = 3000\ kg,\ y$$

$$N\left(n.\frac{\sigma}{\sqrt{n}}\right) = N\left(\sigma.\sqrt{n}\right) = 100*0.3*\sqrt{20} = 134.16\ kg.$$

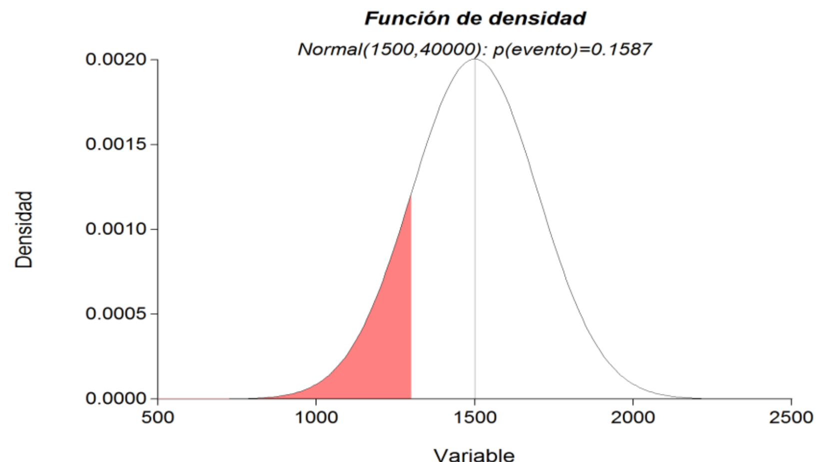
$\hat{\tau} \sim N(3000\,, 134.16)$. La varianza es igual a: 134.16^2=18,000.

El valor de probabilidad requerido es P(X<3300 kg). **R/ 0.9873 (98.73%)**

Ejemplo 73

Los pesos de los pollos de una granja se distribuyen normalmente con media de 1500 gr. y una desviación estándar de 200 gr. (varianza de 40000 u^2).

a) Si se rechazan los pollos que pesan menos de 1300 gr. ¿qué tanto por ciento hay que rechazar?



b) Elegidos dos pollos al azar, ¿cuál es la probabilidad de que por lo menos un pollo pese más de 1800 gr? Resolver utilizando la distribución binomial, con n=2 y **p= P(X>1800)= 0.0668073** y calcular la probabilidad de que **P(X≥1)= P(X>0) = 0.12913776.**

Calculador de probabilidades y cuantiles

Seleccione la distribución: Uniforme (a,b); Normal (media,varianza); T Student (v); Chi Cuadrado (v,lambda); F no central (u,v,lambda); Exponencial (lambda); Gama (lambda,r); Beta (a,b); Weibull (a,b); Doble Exponencial (a,b); Logística (a,b); Pareto (Theta,a); Gumbel (a,b); Rangos estudent.(k,v); Poisson (lambda); Binomial (n,p); Geométrica (p); Hipergeométrica (m,k,n); Binomial Negativa (m,k); Beta-Binomial(p,Rho,N)

1500 media
40000 varianza
Valor de x: 1800
Prob. (X<=x): 0.9331926954
Prob. (X>x): 0.06680730458
Prob. (X=x): 0
Calcular | Ayuda

Calculador de probabilidades y cuantiles

Seleccione la distribución: Uniforme (a,b); Normal (media,varianza); T Student (v); Chi Cuadrado (v,lambda); F no central (u,v,lambda); Exponencial (lambda); Gama (lambda,r); Beta (a,b); Weibull (a,b); Doble Exponencial (a,b); Logística (a,b); Pareto (Theta,a); Gumbel (a,b); Rangos estudent.(k,v); Poisson (lambda); Binomial (n,p); Geométrica (p); Hipergeométrica (m,k,n); Binomial Negativa (m,k); Beta-Binomial(p,Rho,N)

2 n
0.0668 p
Valor de x: 0
Prob. (X<=x): 0.87086224
Prob. (X>x): 0.12913776
Prob. (X=x): 0.87086224
Calcular | Ayuda

LISTA DE EJERCICIOS 7

1. Un ejecutivo de una cadena de televisión está estudiando propuestas para nuevas series. A su juicio, la probabilidad de que una serie tenga una audiencia mayor que 17.8 es 0.25, además, probabilidad de que la serie tenga una audiencia mayor que 19.2 es 0.15. Si la incertidumbre de este ejecutivo puede representarse mediante una variable aleatoria normal. ¿Cuál es la media y la desviación estándar de esta distribución?

2. Estudios meteorológicos indican que la precipitación pluvial mensual en períodos de sequía es una cierta región puede ser considerada como una variable aleatoria que sigue aproximadamente una distribución normal, con media de 30 mm y varianza de 16 mm^2.

a) ¿Cuál sería el valor de precipitación pluvial de modo que existe apenas 10% de probabilidad de haber una precipitación inferior a ese valor?

b) Construya un intervalo central que contenga 80% de los posibles valores de precipitación pluvial

c) Admitiendo que ese modelo es correcto para los próximos 50 meses, ¿en cuántos de ellos esperaríamos una precipitación superior a 34 mm?

3. La distribución de los pesos de conejos criados en una granja puede ser bien representada por una distribución normal, con media de 5 kg y desviación estándar de 0.8 kg. Una empresa dedicada a la comercialización de carne, comprará 5,000 conejos y pretende clasificarlos de acuerdo con el peso, de la siguiente manera: el 20% de los menos pesados como PEQUEÑOS, los 55% siguientes como MEDIANOS, los 15% siguientes como GRANDES, y el 10% más pesados como EXTRAGRANDES. Determine los límites de clase para cada una de las categorías.

4. La vida útil de cierto tipo de lavadora automática sigue aproximadamente una distribución normal, con media y desviación estándar de 3.1 y 1.2 años respectivamente. Si este tipo de lavadora tiene garantía de un año ¿Qué fracción de la cantidad vendida originalmente necesitará ser reemplazada?

5. El tiempo promedio requerido para terminar un examen es de 70 minutos, con desviación estándar de 12 minutos. ¿Cuánto tiempo debe asignarse si se desea que el 85% de los estudiantes termine el examen?

6. Una variable X tiene una distribución normal con una media μ desconocida y una desviación estándar $\sigma=2$. Si la probabilidad de que x sea mayor que 7.5 es 0.8, encuentre μ.

7. Una empresa formuladora de agroquímicos produce un compuesto químico y está preocupada por su contenido de impurezas. Se estima que el peso de las impurezas por lote se distribuye según una distribución normal con media de 12.2 gramos y desviación estándar de 2.8 gramos. Si se elige un lote al azar:

a) ¿Cuál es la probabilidad de que contenga menos de 10 gramos de impurezas?
b) ¿Cuál es la probabilidad de que contenga más de 15 gramos de impurezas?
c) ¿Cuál es la probabilidad de que contenga entre 12 y 15 gramos de impurezas?

8. Sea Z una variable aleatoria normal estándar, encuentre:

a) $P(Z<1.20)$
b) $P(Z>1.33)$
c) $P(Z< -1.70)$
d) $P(Z> -1.0)$
e) $P(1.20<Z< 1.33)$
f) $P(-1.70 < Z < -1.0)$

9. El 31% de los alumnos están bajo 45 puntos en una prueba y un 8% sobre 64 puntos. ¿Cuál es la media y la desviación estándar de la prueba?

10. La estatura media de los hombres adultos es de 170 cm. El 10% de ellos mide más de 175 cm, suponiendo que la estatura de los alumnos del curso es normal. ¿Cuál es la desviación estándar?

11. La estatura media de los estudiantes está distribuida normalmente. Si el 13.57% de los estudiantes mide más de 174.4 cm y el 8.08% mide menos de 164.4 cm ¿Cuál es la media y la desviación estándar de la estatura de los estudiantes?

12. A un bosque de *Pinus taeda* se le midió el DAP (diámetro a la altura del pecho) que sigue una distribución normal con media de 21cm y desviación estándar de 5 cm.

a) Si todos los árboles con DAP < 18cm fueran cortados, ¿cuál es la proporción de árboles cortados?

b) Sí un fitomejorador forestal selecciona al 2.5% de los árboles mayores del bosque, ¿cuál es el DAP mínimo de los árboles seleccionados?

c) Un Ingeniero Forestal desea cortar 20% de los árboles a partir de los menores diámetros. ¿Cuál es el DAP máximo de los árboles a ser cortados?

d) Un aserradero requerirá árboles con DAP entre 18 e 27 cm. ¿Cuál es la proporción de árboles de este bosque que podrán ser utilizados en este aserradero?

13. Los naranjos de una plantación ubicada en Coatepeque, tienen una producción que en peso se distribuye normalmente. El 25 % de los árboles tienen más de 51 kg de fruta y el 60 % más de 40 kg. ¿Cuál es la media y la desviación estándar de la distribución?

14. Una empresa produce 1 millón de botellas de jugo de naranja al mes, cuyos pesos siguen una distribución normal con media de 1,200 g. y desviación estándar de 10 g.. Calcule para un mes:

a) El número de botellas que pesan más de 1,225 g.
b) El número de botellas que pesan entre 1,195 g. y 1,215 g.
c) El número de botellas que pesan menos de 1,190 g.

15. La distribución de los diámetros de una plantación de *Pinus caribaea* var. *caribaea* sigue una distribución normal con media 23 cm. y desviación estándar de 7 cm. Con esta información calcule:

a) ¿Cuál es el diámetro mínimo para que un árbol esté entre el 1% de los mayores?
b) ¿Cuál es la proporción de árboles con diámetros entre 16 y 30 cm?
c) ¿Cuál es la proporción de árboles con diámetros arriba de 20cm?

16. En una plantación forestal la proporción de árboles con la altura arriba de 4 metros es 0.80 y arriba de 7 metros es de 0.30. En 3000 árboles, ¿cuántos tendrán altura de 5.5 metros o más?

17. Un granja provee pollos en cajas de 20 animales, con peso medio de 1.5 kg y desviación estándar de 0.3 kg. El peso de cada caja es constante e igual a 2 kg. En un cargamento de 100 cajas, cuál es la probabilidad de que el peso total:

    e) sea inferior a 3200 kg.,
    f) esté entre 2800 y 3000 kg.,
    g) sea superior a 2700 kg.

18. En una lechería la producción de leche por vaca tiene distribución normal con media igual a 18 litros y varianza de 9 u^2. ¿Cuál es la probabilidad que una vaca elegida al azar:

    a) produzca menos de 12 litros?
    b) tenga una producción entre 21 y 24 litros?
    c) produzca entre 15 y 22 litros?
    h) tenga una producción mayor de 25 litros?
    i) si la lechería cuenta con 3000 vacas la pregunta de cuántas de ellas producen entre 15 y 21 litros
    j) ¿cuántas vacas producirán más de 24 litros?

19. Si en un cierto huerto de Chichicastenango el peso de manzanas Red Delicious, tiene distribución normal con media 140 gr y desviación estándar de 20 gr, determine:

    a) el peso máximo del 10% de las manzanas de menor peso, o sea, el percentil 10.
    b) el peso mínimo del 5% de las manzanas más grandes, es decir, el percentil 95.
    c) entre que peso se encuentra el 90% central de las manzanas.

20. Utilice en Infostat el menú ESTADISTICA ----- PROBABILIDADES y CUANTILES para resolver los siguientes cuestionamientos:

    a) ¿Cuál es el valor de z, tal que, $P(x \geq z) = 0.95$?
    b) ¿Cuál es el valor de z, tal que, $P(x \leq z) = 0.90$?

21. En un matadero la experiencia demuestra que el peso de los terneros que se sacrifican sigue una distribución normal con media de 150 kg y desviación estándar de 10 kg:

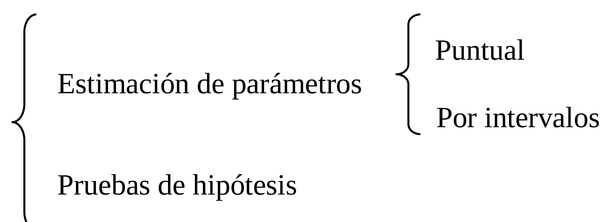
    a) Si se selecciona un ternero al azar y lo pesamos, ¿cuál es la probabilidad de que pese más de 160 kg?
    b) Si se seleccionan tres terneros al azar y lo pesamos, ¿cuál es la probabilidad de que los tres pesen más de 160 kg?

22. Se tiene para la venta un lote de 1000 pollos, con un peso promedio de 3.56 kg y una desviación estándar de 0.18 kg. ¿Cuál es la probabilidad que una muestra aleatoria de 100 pollos extraída de esa población pesen entre 3.53 y 3.56 kg?

UNIDAD IV
ESTIMACIÓN

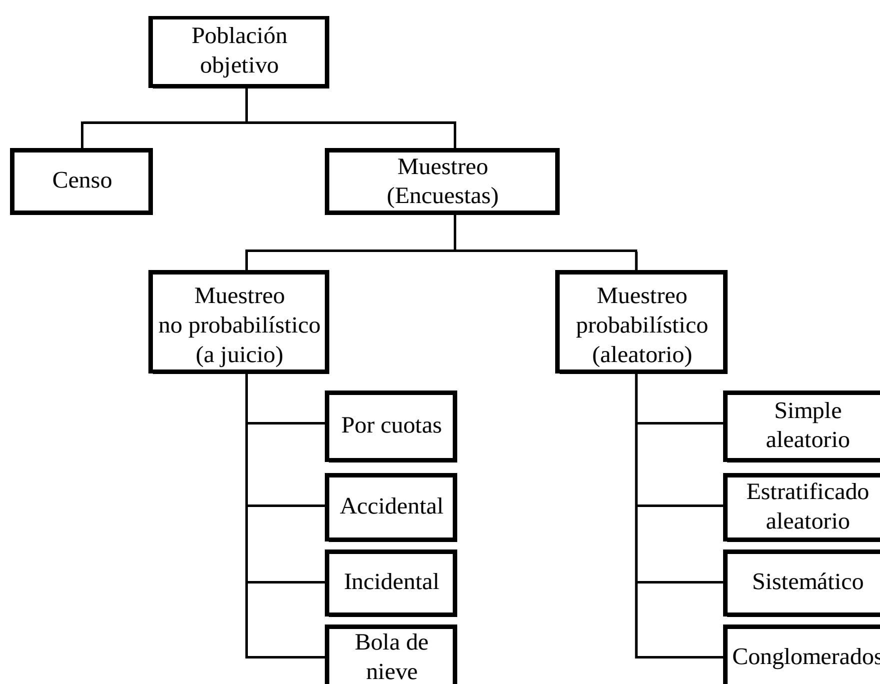
4.1 INFERENCIA ESTADÍSTICA

La Inferencia es la rama de la estadística que tiene por objeto estudiar la población a través de evidencias proporcionadas por la muestra. La inferencia puede ser definida de la siguiente forma:



Uno de los principales problemas que se presentan en la estadística es el de hacer afirmaciones sobre los parámetros poblacionales (generalmente desconocidos), por ejemplo, saber cuál es el tiempo necesario para que el organismo humano pueda degradar cierto compuesto químico, cuál es la producción total de maíz blanco en Guatemala en el año 2007, cuál es la altura media de la población masculina guatemalteca. Y para responder a estas preguntas, muchas veces tenemos que hacer uso del proceso de muestreo, que consiste en estudiar apenas una fracción de la población (una muestra) y a partir de ésta, hacer inferencias sobre la población.

Para que el proceso anteriormente descrito sea confiable, es necesario que la muestra utilizada sea representativa de la población, y para eso, ella debe ser retirada según determinadas técnicas de muestreo. Los tipos principales de muestreo se resumen en el siguiente esquema:



Para hacer inferencias sobre los parámetros poblacionales de esta muestra, es necesario el conocimiento de las relaciones existentes entre las estimativas obtenidas y los valores de los parámetros poblacionales, o sea, es necesario conocer la distribución muestral del estimador utilizado, para que se pueda hacer una inferencia segura sobre un parámetro cualquiera.

4.2 DISTRIBUCIONES DE MUESTREO

Una vez seleccionada la muestra a estudiar se calculan los estadísticos con el fin de estimar los parámetros poblacionales (por ejemplo: media aritmética, proporción, total, desviación estándar, etc.), esto genera un error, el cual se define como **error de muestreo** y corresponde a la diferencia entre el parámetro poblacional y el estimador, e indica la precisión del estimador.

Debido al error de muestreo cuando se utilizan los resultados muestrales para calcular un parámetro poblacional surge la siguiente pregunta:

¿Cómo podemos realizar una predicción acertada acerca de la población usando datos de la muestra?

a) Conociendo la distribución de muestreo de la estadística
b) Haciendo uso del Teorema Central del Límite

La distribución de todos los valores que puede asumir una estadística, calculados a partir de un número grande de muestras del mismo tamaño, seleccionadas en forma aleatoria de la misma población se llama: **Distribución muestral de esa estadística.**

IMPORTANTE

Cualquier distribución de probabilidad (y por lo tanto, cualquier distribución de muestreo) puede ser descrita parcialmente por su media y su desviación estándar. La media en este caso, estará dada por la media de las medias de todas las muestras, y que es equivalente a la media poblacional.

Por otra parte, la desviación estándar de la distribución de las medias de las muestras mide el grado hasta el cual esperamos que varíen las diferentes muestras debido al error de muestreo. La desviación estándar de la distribución de muestreo de una estadística se le conoce como: **ERROR ESTANDAR DE LA ESTADÍSTICA,** se denota con el símbolo ($\sigma_{\bar{x}}$) y se calcula usando la siguiente expresión: $\sigma / \sqrt{n}$, siendo σ la desviación estándar poblacional y n el tamaño de la muestra.

Multiplicador de población finita para el error estándar de la media ($\sigma_{\bar{x}}$)

Cuando la población es finita y el muestreo se realiza sin reemplazo, $\sigma_{\bar{x}}$ es dado por:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Multiplicador de población finita

Nota: Se debe utilizar el multiplicador de población finita para corregir el error estándar de la media, cuando la fracción de muestreo n /N > 0.05.

Distribución de muestral de $\bar{x}$ a partir de poblaciones que no siguen una distribución normal.

Cuando el muestreo se efectúa a partir de una población que no sigue una distribución normal, se utiliza un teorema matemático conocido como: TEOREMA CENTRAL DEL LIMITE. Este teorema se cita a continuación:

4.3 TEOREMA CENTRAL DEL LÍMITE

El teorema central del límite permite tomar muestras a partir de poblaciones con distribución no normal y garantizar que se obtengan aproximadamente los mismos resultados que si la población tuviera una distribución normal, siempre que se tome una muestra grande.

4.4 DISTRIBUCIÓN MUESTRAL DE MEDIAS Y DE PROPORCIONES.

4.4.1 DISTRIBUCIÓN DE MUESTREO DE LA MEDIA MUESTRAL

Recuerde que:

Si se selecciona una muestra aleatoria de n mediciones de una población con media μ y desviación estándar σ, la distribución de muestreo de la media muestral $\bar{x}$ tendrá las siguientes propiedades:

Media: $\mu_{\bar{x}} = \mu$ (o sea, la media de todas las medias muestrales es igual a la media poblacional).

Desviación estándar: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Conocida como error estándar y es igual a la desviación estándar de la población dividida entre la raíz cuadrada del tamaño de la muestra.

Ejemplo 74

Una institución bancaria calcula que sus cuentas de ahorro individuales están normalmente distribuidas con una media de US$ 2,000 y una desviación estándar (σ) de US$ 600. Si el banco toma una muestra aleatoria (n) de 100 cuentas, ¿cuál es la probabilidad de que la media de la muestra caiga entre US$ 1,900 y US$ 2,050?

Solución:

1. Ésta es una pregunta con respecto a la distribución de muestreo de la media, por tanto es necesario calcular primero el error estándar de la media.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{US\$600}{\sqrt{100}} = \frac{US\$600}{10} = US\$60$$

2. Para calcular la probabilidad de que la media de la muestra caiga entre US$ 1,900 y US$ 2,050, se utiliza la tabla de valores de z y la siguiente ecuación:

Media de la muestra → $\bar{x}$; Media de la población → μ ; Error estándar de la media → $\sigma_{\bar{x}}$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

2.1 Para $\bar{x}$ = US$ 1,900, se tiene que:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{1{,}900 - 2{,}000}{60} = -\frac{100}{60} = -1.67$$

El –1.67, indica el número de desviaciones estándar que hay de distancia entre el valor de la media poblacional y el valor de la media muestral.

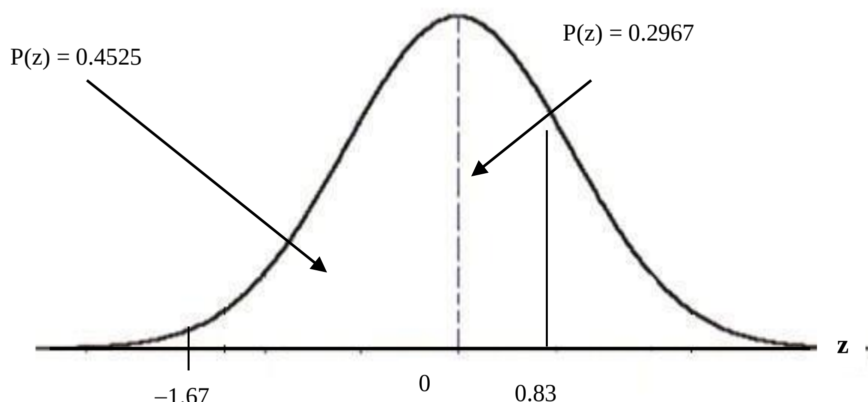
2.2 Para $\bar{x}$ = US$ 2,050, se tiene que:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{2{,}050 - 2{,}000}{60} = -\frac{50}{60} = 0.83$$

De acuerdo con la tabla de valores de z, para el valor z = –1.67 (buscamos como valor absoluto, o sea 1.67) corresponden un área de 0.4525, y para z =0.83 el área es de 0.2967. A continuación se ilustra cómo obtener esos valores en la tabla de z.

Z	0	1	2	3	4	5	6	7	8	9
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670

Estas áreas se representan en el gráfico de la distribución normal, que tiene forma de campana.



Si sumamos estos valores: 0.4525, que es el área comprendida entre z =0 y z = –1.67 y 0.2967, el área comprendida entre z =0 y z= 0.83, obtenemos 0.7492 como la probabilidad total de que media de la muestra se encuentre entre US$ 1,900 y US$ 2,050.

Ejemplo 75

En una empresa de alimentos, una máquina empaca cajas con cereal, y está ajustada de modo que la cantidad de cereal en una caja sea normalmente distribuida con una media aritmética de 368 gramos. A partir de experiencias anteriores, la desviación estándar poblacional para este proceso es conocida y es igual a 15 gramos. Si una muestra de 25 cajas es seleccionada aleatoriamente de las miles que son producidas diariamente, y se obtiene el peso promedio de esta muestra, obtenga la probabilidad de que:

a) El peso promedio de la muestra este entre 370 y 373 gramos.
b) El peso promedio sea menor a 365 gramos.
c) El peso promedio esté entre 365 y 370 gramos.

Solución:

1. Ésta es una pregunta con respecto a la distribución de muestreo de la media, por tanto es necesario calcular primero el error estándar de la media.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$$

2. Para calcular la probabilidad de que la media de la muestra esté entre 370 y 373 gramos, se utiliza la tabla de valores de z, por lo que hay que obtener previamente estos valores, tal como fue mostrado en el Caso No. 7

2.1 Para $\bar{x}$ = 373, se tiene que:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{373 - 368}{3} = \frac{5}{3} = 1.666$$

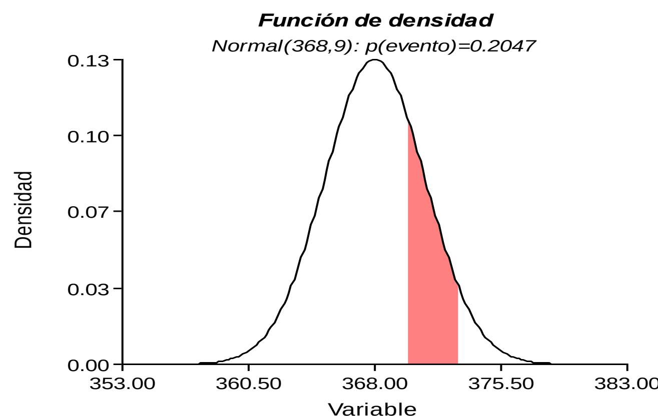
2.2 Para $\bar{x}$ = 370, se tiene que:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{370 - 368}{3} = \frac{2}{3} = 0.666$$

De acuerdo con la tabla de valores de z, para el valor z = 1.66 corresponden un área de 0.4515, y para z =0.66 el área es de 0.2454. A continuación se ilustra cómo obtener esos valores en la tabla de z.

Z	0	1	2	3	4	5	6
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154

Estas áreas se representan en el gráfico de la distribución normal, que tiene forma de campana

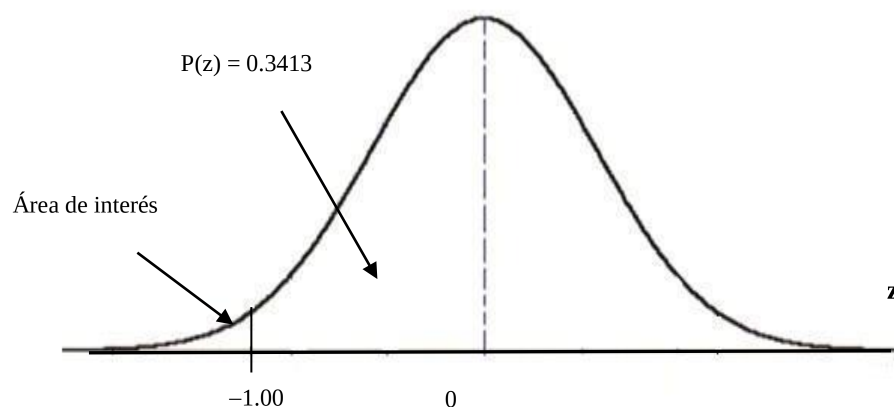


El área de 0.4515 va desde z=0 hasta z= 1.66 y equivale a 0.4515; por otra parte, el área de 0.2454 va desde z= 0 hasta z= 0.66, por lo tanto si deseamos obtener la probabilidad de que la media de la muestra esté entre 370 y 373 gramos, debemos restar ambas áreas y obtenemos el área de interés que se muestra en la curva normal. Por tanto 0.4515–0.2454 = 0.2061 es la probabilidad de que media de la muestra se encuentre entre 370 y 373 gramos.

3. Para calcular la probabilidad de que la media de la muestra sea menor a 365 gramos, se obtiene el valor de z correspondiente:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{365 - 368}{3} = \frac{-3}{3} = -1.0$$

De acuerdo con la tabla de valores de z, para el valor z = –1.0 corresponden un área de 0.3413. Este valor de área va desde z=0 hasta z= –1.0. Pero como nuestra área de interés son los valores <360, tenemos que restar a 0.5 (que la mitad del área total bajo la curva, recuerde que la distribución normal es simétrica y que el área total bajo la curva es igual a 1) el valor 0.3413, y obtenemos 0.1587, la probabilidad de que la media de la muestra sea menor a 360 gramos.

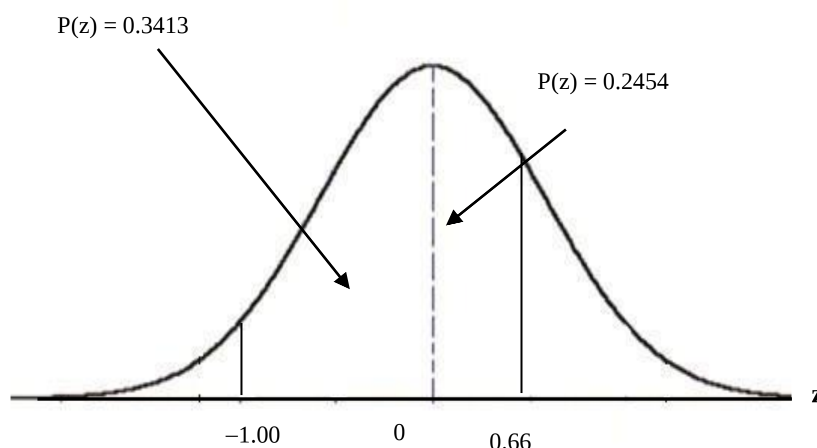


4. Para calcular la probabilidad de que la media de la muestra esté entre 360 y 370 gramos, tomemos los resultados obtenidos en los incisos anteriores:

Para $\bar{x}$ = 370: $$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{370 - 368}{3} = \frac{2}{3} = 0.666$$ P(z) = 0.2454

Para $\bar{x}$ = 365: $$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{365 - 368}{3} = \frac{-3}{3} = -1.0$$ P(z) = 0.3413

Gráficamente podemos representar estos valores, así:



Por lo tanto la probabilidad de que la media de la muestra esté entre 360 y 370 gramos, es igual a la suma de ambas áreas, 0.3413 + 0.2454 = 0.5867

Ejemplo 76

Juan Luís Pérez, auditor de la compañía de tarjetas de crédito MAYACARD, sabe que el saldo promedio mensual de un cliente dado es de $ 112 y la desviación estándar es de $ 56. Si Juan audita 50 cuentas seleccionadas al azar, encuentre la probabilidad de que el saldo promedio mensual de la muestra sea:

a) Menor que $ 100.
b) Entre $ 100 y $ 130.

4.4.2 DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

En muchos casos de los negocios y la economía se usa proporción muestral p para hacer inferencias estadísticas sobre la proporción p.

Definición de proporción

Si se selecciona una muestra aleatoria de n personas de la población y si x de ellas tienen la característica de interés, entonces se utiliza la proporción muestral:

$$p = \frac{x}{n},$$

para estimar proporción poblacional p.

La distribución de muestreo de la media muestral $\bar{x}$ tendrá las siguientes propiedades:

Media: $\mu_p = p$ (o sea, la media de todas las medias muestrales es igual a la media poblacional).

Desviación estándar: $\sigma_p = \sqrt{\frac{p\,(1-p)}{n}}$ o $\sigma_p = \sqrt{\frac{p\,(1-p)}{n}}\sqrt{\frac{N-n}{N}}$

Ejemplo 77

Una muestra aleatoria de n = 100 mayoristas que compran tubos de plástico PVC indicó que 59 planean incrementar sus compras el próximo año. Estime la proporción p de los comerciantes en la población total de mayoristas de dicho material que planean incrementar sus compras el año próximo, y obtenga además, el error estándar para esa proporción.

Solución:

$p = \frac{x}{n} = \frac{59}{100} = 0.59$ y $\sigma_p = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{0.59\,(1-0.59)}{100}} = 0.049$

4.5 ESTIMACIÓN

En estadística se llama estimación al conjunto de técnicas que permiten dar un valor aproximado de un parámetro de una población a partir de los datos proporcionados por una muestra. Es necesario indicar que el término estimación también se utiliza en ciencias aplicadas para hacer referencia a un cálculo aproximado, que normalmente se apoya en la herramienta estadística aunque puede no hacerlo.

4.5.1 ESTIMACIÓN PUNTUAL Y ESTIMACIÓN POR INTERVALOS

Podemos hacer dos tipos de estimaciones concernientes a una población: una estimación puntual y una estimación por intervalos.

Una **estimación puntual** es un solo número que se utiliza para estimar un parámetro de población desconocido. Por ejemplo, deseamos saber cuál es el consumo promedio de energía eléctrica de los habitantes del barrio Candelaria, municipio de Coatepeque, Quetzaltenango. Luego de realizar un muestreo a un grupo de hogares, un estudiante del curso de Estadística determinó que el consumo promedio mensual para el cuarto trimestre del año 2007 sería de 105 kWh. Esta es una estimación puntual, ya que representa un único valor.

Un procedimiento de estimación puntual utiliza la información de una muestra para llegar a un solo número, o punto, que estima el parámetro de interés. La estimación real se realiza mediante un **estimador**.

Definición: Un estimador es una regla que expresa cómo calcular la estimación, basándose en la información de la muestra y se enuncia, en general, mediante una ecuación.

Por ejemplo, la media muestral: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Es un estimador puntual de la media poblacional μ y explica exactamente cómo puede obtenerse el valor numérico de la estimación, una vez conocidos los valores muestrales $x_1, x_2, \ldots, x_n$.

Importante:

Una estimación puntual a menudo resulta insuficiente, debido a que sólo tiene dos opciones: es correcta o está equivocada. Si se nos dice solamente que la afirmación del estudiante del curso de Estadística está equivocada, usted no sabe qué tan mal está, y no puede tener la certeza de la confiabilidad de la estimación. En consecuencia, una estimación puntual es mucho más útil si viene acompañada por una estimación del error que podría estar implicado.

Una **estimación por intervalo** es un intervalo de valores que se utiliza para estimar un parámetro de población. Esta estimación indica el error de dos maneras: por la extensión del intervalo y por la probabilidad de obtener el verdadero parámetro de la población que se encuentra dentro del intervalo. Para el ejemplo anterior del consumo de energía eléctrica, el estudiante podría llegar a la siguiente conclusión: estimo que el consumo promedio mensual de energía eléctrica por hogar en el barrio Candelaria para el cuarto trimestre del año 2007 estará entre 100 y 110 kWh.

El intervalo de confianza describe la variabilidad entre la medida obtenida en un estudio y la medida real de la población (parámetro). Corresponde a un rango de valores, cuya distribución es normal y en el cual

se encuentra, con alta probabilidad, el valor real de una determinada variable. Esta «alta probabilidad» se ha establecido por consenso en 95%. Así, un intervalo de confianza de 95% nos indica que dentro del rango dado se encuentra el valor real de un parámetro con 95% de certeza

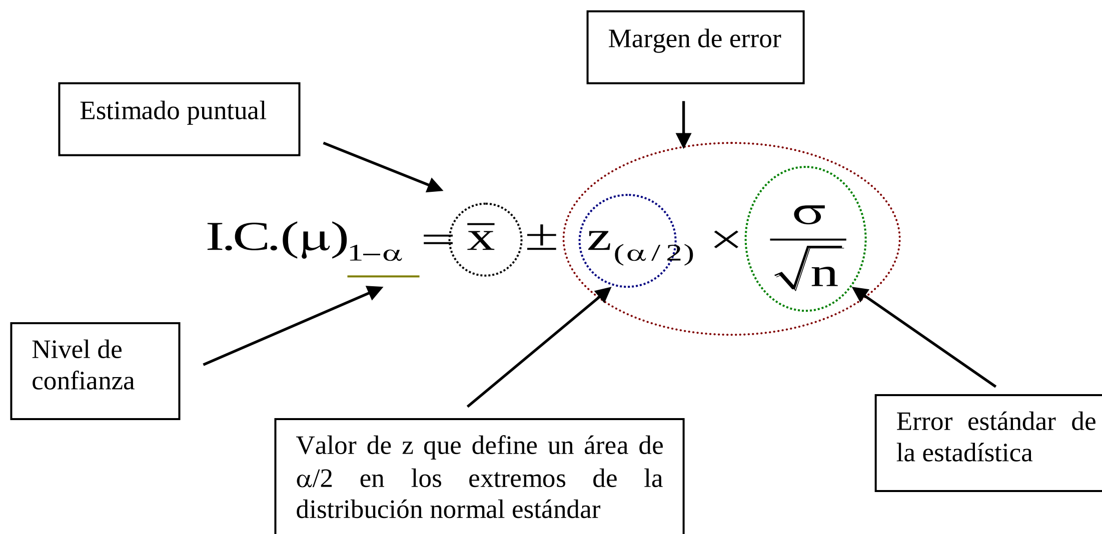
En resumen, la estimación por intervalos utiliza los datos de una muestra para determinar dos puntos que pretenden abarcar el valor real del parámetro estimado. Por lo que hay que advertir sobre la relación inversa entre la amplitud del IC y el tamaño muestral, mientras más pequeño sea el *n* más amplio es el IC. A mayor número tamaño de *n*, más certeza tenemos que el resultado del experimento se acerca al valor real, por lo tanto el IC es más estrecho.

Definición de nivel de confianza

En estadística, la probabilidad que asociamos con una estimación de intervalo se conoce como nivel de confianza. Esta probabilidad indica qué tanta confianza tenemos de que la estimación de intervalo incluya al parámetro de población. Una probabilidad más alta significa más confianza. En estimación, los niveles de confianza más utilizados son: 90%, 95% y 99%, pero somos libres de aplicar cualquier nivel de confianza.

Estructura de un intervalo de confianza

Para ejemplificar la estructura de un intervalo de confianza, vamos a suponer que fue sorteada una muestra de tamaño *n*, encontrada su media $\bar{x}$ y suponiendo la varianza poblacional es conocida σ^2, podemos construir el intervalo de confianza (IC) para la media poblacional μ, así:



Características de un buen estimador

Antes de emplear algún estadístico de muestra como estimador puntual, se debe comprobar si tiene ciertas propiedades asociadas a los buenos estimadores puntuales: imparcialidad (in sesgo), eficiencia, consistencia y suficiencia.

Imparcialidad: se refiere al hecho de que una media de una muestra es un estimador no sesgado de una media de población porque la media de la distribución de muestreo de todas las medias muestrales tomadas de la misma población es igual a la media de la población misma.

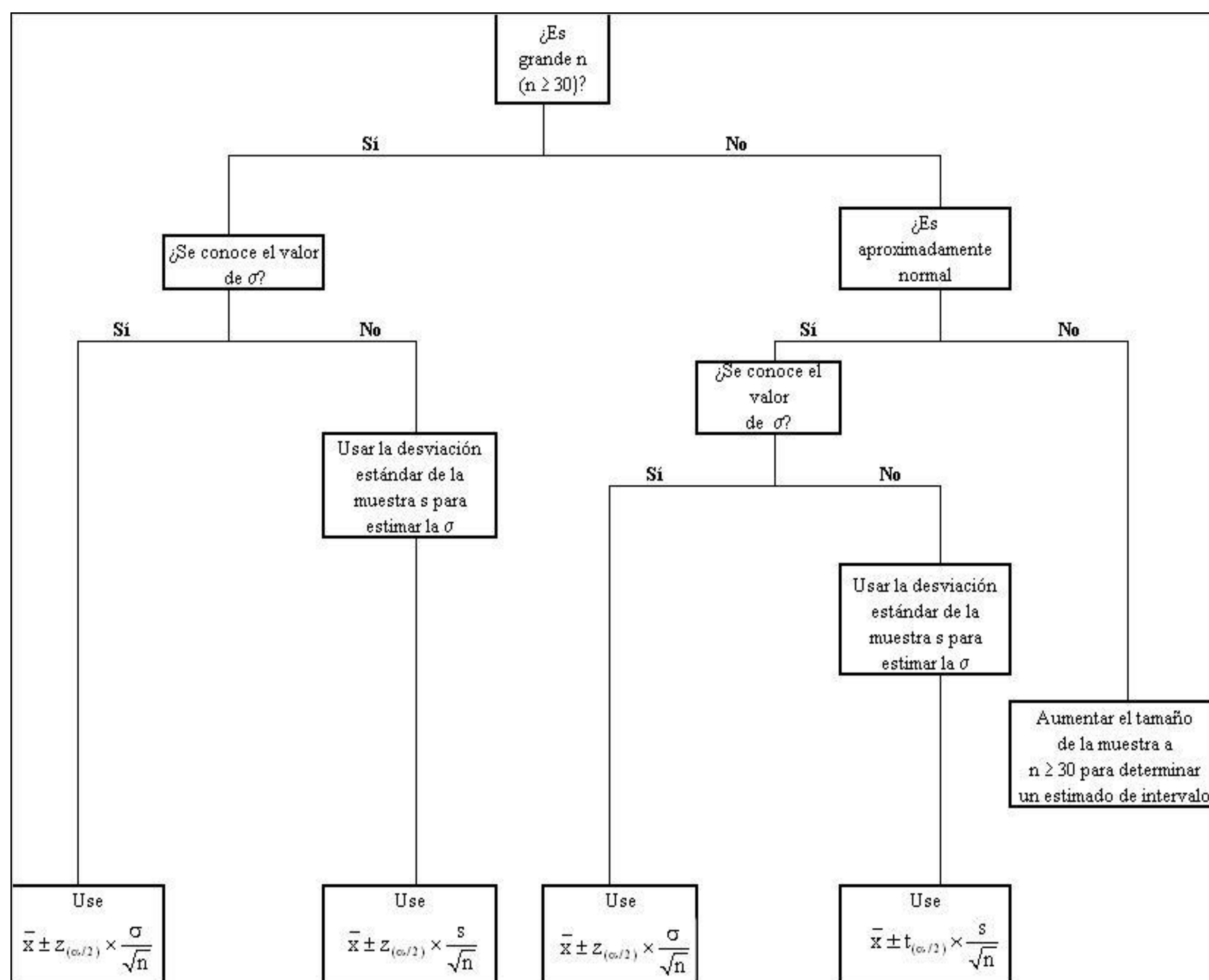
Eficiencia: se refiere a la precisión de la estadística de muestreo como un medio para estimar el parámetro de la población, lo cual se ve afectado por el tamaño del error estándar de la estadística. Si comparamos dos estadísticas de una muestra del mismo tamaño y tratamos de decidir cuál de ellas es un estimador más eficiente, escogeríamos la estadística que tuviera el menor error estándar o la menor desviación estándar de la distribución de muestreo (**mínima varianza**)

Consistencia: un estimador puntual es consistente si sus valores tienden a acercarse al parámetro de población conforme se incrementa el tamaño de la muestra. En otras palabras, un tamaño grande de muestra tiende a proporcionar un mejor estimador puntual que un tamaño pequeño.

Suficiencia: un estimador es suficiente si utiliza una cantidad de información contenida en la muestra que ningún otro estimador podría extraer información adicional de la muestra sobre el parámetro de la población que se está estimando.

4.5.2 ESTIMACIÓN PUNTUAL Y DE INTERVALO PARA UNA MEDIA POBLACIONAL

En el siguiente esquema se presenta un resumen de procedimientos de estimación por intervalos para una media poblacional.



Ejemplo 78 Población normal (n>30), varianza poblacional σ^2 conocida.

Un vendedor al mayoreo de partes automotrices necesita una estimación de la duración media que puede esperar de los limpiabrisas en condiciones normales de manejo. La administración de la empresa ya ha determinado que la desviación estándar (σ) de la vida útil de la población es de seis meses. Suponga que seleccionamos una sola muestra aleatoria de n = 100 limpiadores, tomamos los datos referentes a su vida útil y obtenemos que la media de esta muestra $\bar{x} = 21$ meses.

Como el vendedor utiliza 10,000 de estos limpiabrisas al año, nos pide que encontremos una estimación de intervalo con un nivel de 95% de confianza.

Solución:

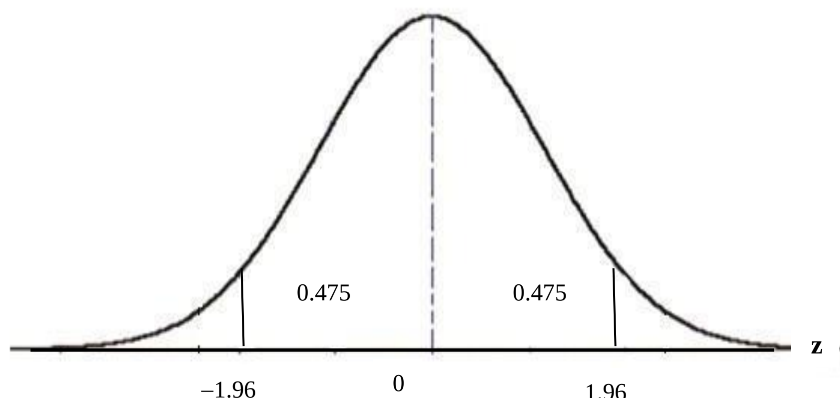
Como el tamaño de la muestra es mayor que 30, de modo que el teorema central del límite nos permite usar la distribución normal como nuestra distribución de muestreo, incluso si nuestra población no está normalmente distribuida. El procedimiento para encontrar los intervalos de confianza se resume en los pasos siguientes:

1. Calculamos el error estándar de la media para una población infinita:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6 \text{ meses}}{\sqrt{100}} = 0.6 \text{ meses}$$

2. Obtenemos el valor de z de acuerdo con el nivel de confianza definido.

Como un nivel de 95% de confianza incluirá 47.5% del área que se encuentra a ambos lados de la media de la distribución de muestreo, podemos buscar en el cuerpo de la tabla de distribución normal estándar el valor correspondiente a 0.475. Descubrimos que 0.475 del área bajo la curva normal está contenida entre la media y un punto situado a 1.96 errores estándar hacia la derecha de la media. Por consiguiente sabemos que (2) (0.475) = 0.95 del área está localizada entre más menos 1.96 errores estándar de la media.



3. Luego se calculan los límites de confianza.

Límite superior de confianza: $\bar{x} + z_{(\alpha/2)} \times \sigma_{\bar{x}}$

Límite inferior de confianza: $\bar{x} - z_{(\alpha/2)} \times \sigma_{\bar{x}}$

Sustituyendo los valores numéricos en estas dos expresiones, tenemos:

Límite superior de confianza: $21 \text{ meses} + (1.96 \times 0.6 \text{ meses}) = 22.18 \text{ meses}$

Límite inferior de confianza: $21 \text{ meses} - (1.96 \times 0.6 \text{ meses}) = 19.82 \text{ meses}$

4. Conclusión.

Con estos resultados podemos informar que estimamos la vida media de la población de limpiabrisas está entre 19.82 y 22.18 meses con 95% de confianza. Esto se puede expresar así: **19.82 meses ≤ μ ≤ 22.18 meses**

Ejemplo 79 Población normal (n>30), varianza poblacional σ^2 desconocida.

Una compañía de seguros de vida está interesada en estimar el ingreso medio anual de N = 700 familias que viven en un condominio residencial de la ciudad de Guatemala. Para ello se tomó una muestra aleatoria simple de tamaño n = 50 familias, y se encontraron los siguientes resultados: $\bar{x}$ = $ 11,800 y s = $ 950 (desviación estándar de la muestra).

La empresa nos solicita que realicemos una estimación por intervalo del ingreso anual medio de las 700 familias de modo que pueda tener 90% de confianza de que la media de la población se encuentre dentro de ese intervalo.

Solución:

El tamaño de la muestra es mayor que 30, de manera que, de nuevo, el teorema central del límite nos permite usar la distribución normal como la distribución de muestreo. El procedimiento para encontrar los intervalos de confianza se resume en los pasos siguientes:

1. Calculamos el error estándar de la media. Como no conocemos la desviación estándar de la población, se estimará a través de la desviación estándar de la muestra, que a partir de ahora se identificará con el símbolo $\hat{\sigma}$, que se conoce como: sigma sombrero.

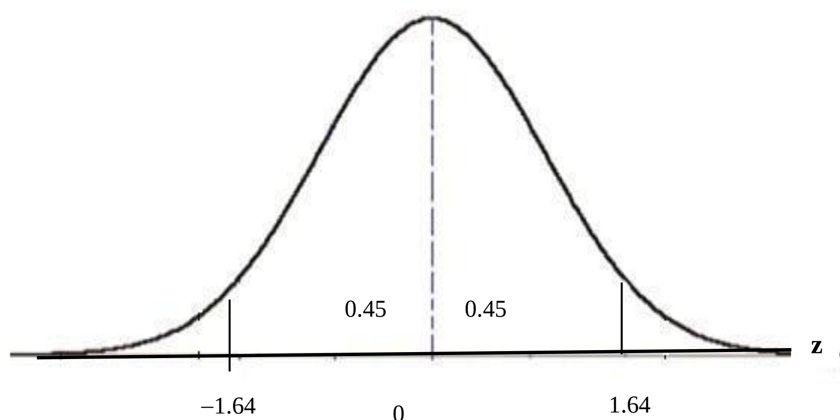
   Ahora ya podemos estimar el error estándar de la media. Como tenemos un tamaño de población finito y puesto que nuestra muestra significa más del 5% de la población (n/N = 50/700 = 0.07 = 7%), utilizaremos la siguiente ecuación:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

$$\hat{\sigma}_{\bar{x}} = \frac{\$\,950}{\sqrt{50}} \times \sqrt{\frac{700-50}{700-1}} = \frac{\$\,950}{7.07} \times \sqrt{\frac{650}{699}} = \$129.57$$

2. Obtenemos el valor de z de acuerdo con el nivel de confianza definido.

El nivel de 90% de confianza incluirá 45% del área que se encuentra a ambos lados de la media de la distribución de muestreo. Si observamos en la tabla de distribución normal estándar el valor correspondiente a 0.45, descubrimos que 0.45 del área bajo la curva normal está contenida entre la media y un punto situado a 1.64 errores estándar hacia la derecha de la media. En consecuencia, 90% del área está localizada entre más menos 1.64 errores estándar de la media.



Z	0	1	2	3	4	5	6	7	8	9
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670

3. Luego se calculan los límites de confianza.

Límite superior de confianza: $\bar{x} + z_{(\alpha/2)} \times \hat{\sigma}_{\bar{x}}$

Límite inferior de confianza: $\bar{x} - z_{(\alpha/2)} \times \hat{\sigma}_{\bar{x}}$

Sustituyendo los valores numéricos en estas dos expresiones, tenemos:

Límite superior de confianza: $\$11,800 + (1.64 \times \$129.57) = \$12,012.50$

Límite inferior de confianza: $\$11,800 - (1.64 \times \$129.57) = \$11,587.50$

4. Conclusión.
El informe que se daría a la Compañía de seguros de vida sería: con un 90% de confianza estimamos que el ingreso anual promedio de las 700 familias que viven en el condominio residencial se encuentra entre $ 11,587.50 y $ 12,012.50.

UN POCO DE HISTORIA

DISTRIBUCIÓN t DE STUDENT

La distribución t de Student fue desarrollada y publicada en 1908 por un investigador de nombre William Sealy Gosset (1876 – 1936)

Gosset trabajaba en la cervecería Guinness en Irlanda y estaba conciente que los propietarios no querían que sus trabajadores publicaran información sobre las investigaciones realizadas en la empresa, talvez por la desconfianza de que secretos industriales fueran de domino público y cayeran en manos de la competencia.

Por eso Gosset, al descubrir una nueva distribución de probabilidades (distribución t) publicó sus trabajos bajo el pseudónimo de Student. Conocedor de las limitaciones que una muestra grande (n > 30) impone al investigador, Gosset creó una estadística adecuada a pequeñas muestras (n ≤ 30). La ecuación de esa estadística es:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$, siendo:

$\bar{x}$ = media aritmética de una muestra pequeña.

μ = media aritmética de la población de donde se extrajo la muestra.

Note que la expresión anterior es parecida a la siguiente ecuación, estudiada en el curso de Estadística I:

$$z = \frac{x_i - \mu}{\sigma}$$. La ecuación para t resulta de la sustitución de x_i por $\bar{x}$ y de σ por $s/\sqrt{n}$

Características de la distribución t de Student:

a) Es simétrica y posee una media = 0.

b) Tiene forma de campana (similar la distribución normal estándar)

c) Varía de $-\infty$ a $+\infty$ (menos infinito a más infinito)

d) No existe una distribución t, sino más bien, una familia de distribuciones t. Para cada n (tamaño de muestra) existe una distribución (y una curva) específica.

e) En la medida que n aumenta, t tiende a z.

Ejemplo 80 Población aproximadamente normal ($n \leq 30$), varianza poblacional σ^2 desconocida.

El director de un hospital privado desea evaluar el tiempo de espera de los pacientes en una clínica. Una muestra aleatoria y representativa de 25 pacientes es seleccionada. El tiempo de espera es definido como el tiempo medido desde la llegada del paciente hasta que es atendido por el médico. Los siguientes datos representan los tiempos de espera (en minutos):

19.5	30.5	45.6	39.8	29.6
25.4	21.8	28.6	52.0	25.4
26.1	31.1	43.1	4.9	12.7
10.7	12.1	1.9	45.9	42.5
41.3	13.8	17.4	39.0	36.6

a) Obtenga una estimación puntual del tiempo promedio de espera.
b) Calcule un intervalo de 95% de confianza para la media poblacional del tiempo de espera, e interprételo.

Solución:

Como el tamaño de la muestra es menor que 30, se utilizará la distribución t de Student. La estimación del intervalo de $(1-\alpha)$ x 100% de confianza para la media poblacional μ con σ desconocida se expresa de la siguiente manera:

$$I.C.(\hat{\mu})_{1-\alpha} = \bar{x} \pm t_{(n-1,\, \alpha/2)} \times \frac{s}{\sqrt{n}}$$, o bien:

$$\bar{x} - t_{(n-1,\, \alpha/2)} \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(n-1,\, \alpha/2)} \times \frac{s}{\sqrt{n}}$$

Siendo $t_{(n-1,\, \alpha/2)}$ el valor crítico de la distribución t, con n–1 grados de libertad, para un área de $\alpha/2$ en la cola superior.

Con los datos de la muestra n = 25, tenemos que $\bar{x} = 27.89$ minutos, s = 13.87 minutos. Para determinar el valor de t de Student, buscamos en la tabla t con 25–1 = 24 grados de libertad y un área $\alpha = 0.05/2 = 0.025$ (área de una cola), ya que utilizaremos 0.95 de confianza, nos queda un área $\alpha/2$ en cada extremo de la distribución. El valor de t se obtiene así:

		Tabla t de Student				
grados de libertad	0.10 **0.20**	0.05 **0.10**	0.025 **0.05**	0.01 **0.02**	0.005 **0.01**	**Una cola** **Dos colas**
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
4	1.533	2.132	2.776	3.747	4.604	
5	1.476	2.015	2.571	3.365	4.032	
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
21	1.323	1.721	2.080	2.518	2.831	
22	1.321	1.717	2.074	2.508	2.819	
23	1.319	1.714	2.069	2.500	2.807	
24	1.318	1.711	2.064	2.492	2.797	

Con esta información se procede a estimar el intervalo de 95% de confianza:

$$I.C.(\mu)_{0.95} = \bar{x} \pm t_{(n-1,\,\alpha/2)} \times \frac{s}{\sqrt{n}} = 27.89 \pm 2.064 \times \frac{13.87}{\sqrt{25}}$$

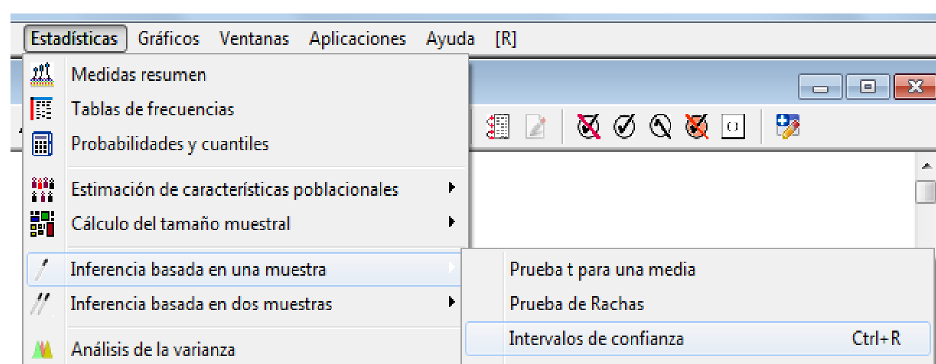
$$= 27.89 \text{ min} \pm 5.73 \text{ min}$$

$$22.17 \text{ min} \leq \mu \leq 33.62 \text{ min}$$

Conclusión:

Podemos concluir con 95% de confianza que el tiempo promedio de espera en la clínica está entre 22.17 y 33.62 minutos. El intervalo de 95% de confianza afirma que tenemos 95% de seguridad de que la muestra que seleccionamos es una muestra en que la media aritmética de la población μ está localizada dentro del intervalo. Esa confianza de 95% significa que, si todas las muestras posibles de tamaño igual a 25 fueran seleccionadas (algo que nunca seria hecho en la práctica), 95% de los intervalos contendrían la verdadera media aritmética de la población, en algún lugar dentro del intervalo.

Ejemplo 81 Cálculo de intervalos de confianza para muestras pequeñas usando Infostat



Intervalos de confianza

Bilateral
Estimación paramétrica

Variable	Parámetro	Estimación	E.E.	n	LI(95%))	LS(95%))
Espera	Media	27.89	2.77	25	22.17	33.62

4.5.3 ESTIMACIÓN PUNTUAL Y DE INTERVALO PARA PROPORCIONES DE UNA POBLACIÓN.

Ejemplo 82

El gerente de producción de un diario desea determinar la proporción de periódicos impresos que presentan algún tipo de problema, tal como exceso de tinta, montaje inapropiado de páginas, falta de páginas, páginas duplicadas, etc. El gerente de producción determinó que fuera seleccionada para análisis una muestra aleatoria de 200 periódicos, en un día cualquiera. Suponga que de esa muestra de 200, 35 contienen algún tipo de problema. Si el gerente de producción desea tener 90% de confianza en la estimación real de la proporción de defectuosos, calcule el intervalo de confianza para esa estimación.

Solución:

La proporción de la muestra p_s = 35/200 = 0.175, con un nivel de confianza de 90%, z = 1.645, por lo que el intervalo se calcula así:

$$p_s \pm z_{(\alpha/2)} \times \sqrt{\frac{p_s\,(1-p_s)}{n}}$$

$$p_s - z_{(\alpha/2)} \times \sqrt{\frac{p_s\,(1-p_s)}{n}} \leq\ p\ \leq p_s + z_{(\alpha/2)} \times \sqrt{\frac{p_s\,(1-p_s)}{n}}$$

$$p_s \pm z_{(\alpha/2)} \times \sqrt{\frac{p_s\,(1-p_s)}{n}}\ =\ 0.175 \pm (1.645)\sqrt{\frac{0.175\,(0.825)}{200}}$$

$$=\ 0.175 \pm (1.645)\,(0.0269)\ =\ 0.175 \pm 0.0442$$

$$0.1308 \leq\ p\ \leq 0.2192$$

Conclusión:

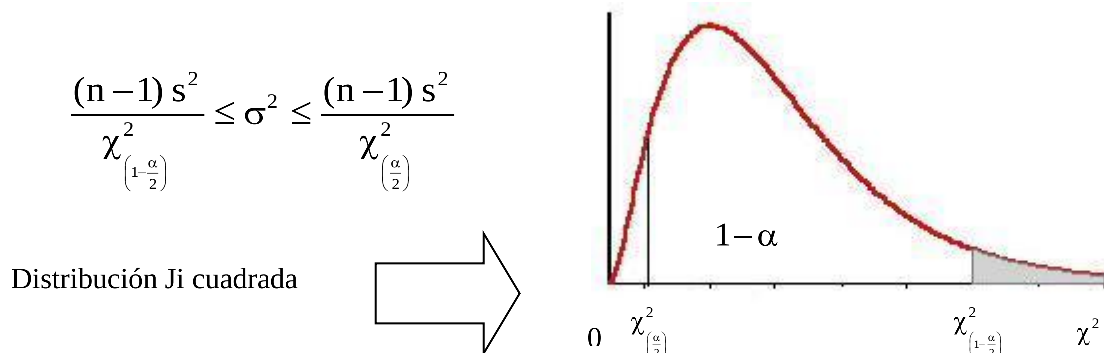
El gerente de producción puede estimar, con 90% de confianza que 13.08% a 21.92% de los periódicos impresos en el día de muestreo presentan algún tipo de problema.

4.5.4 ESTIMACIÓN PUNTUAL Y DE INTERVALO PARA LA VARIANZA DE UNA POBLACIÓN.

Si se toman todas las posibles muestras aleatorias de tamaño n extraídas de una población y se calcula la varianza para cada muestra, se puede obtener la distribución muestral de varianzas.
A cambio de hallar la distribución muestral de s^2 o σ^2 es conveniente hallar la distribución muestral de la variable aleatoria relacionada: $\frac{(n-1)s^2}{\sigma^2}$. Esta variable de muestreo tiene una distribución Ji cuadrada (χ^2) con n–1 grados de libertad. La estimación del intervalo de 100 (1–α)% de confianza para σ^2 se obtiene con las siguientes ecuaciones:

$$\frac{(n-1)\,s^2}{\chi^2_{\left(1-\frac{\alpha}{2}\right)}} \leq \sigma^2 \leq \frac{(n-1)\,s^2}{\chi^2_{\left(\frac{\alpha}{2}\right)}}$$

Distribución Ji cuadrada



Ejemplo 83

Suponiendo que se desee estimar la varianza poblacional para el proceso de llenado de envases con un detergente líquido. Se toma una muestra de 20 envases llenos y se encuentra que la varianza de las cantidades de llenado es s^2= 0.0025 onzas^2. Con esta información, determine un intervalo de 95% de confianza para la varianza poblacional.

Solución:

1– α= 0.95
α = 1– 0.95 = 0.05
s^2 = 0.0025 onzas^2
n = 20
gl = n–1= 19

$\chi^2_{(1-0.05/2)}$ con gl =19 → 32.8523
$\chi^2_{(0.05/2)}$ con gl = 19 → 8.9065

$$\frac{19\times 0.0025}{32.8523} \leq \sigma^2 \leq \frac{19\times 0.0025}{8.90655}$$

$$0.0014 \leq \sigma^2 \leq 0.0053$$

4.5.5 TAMAÑO DE MUESTRA EN FUNCIÓN DEL TAMAÑO DE LA POBLACIÓN Y DEL PARÁMETRO A ESTIMAR.

Para calcular el tamaño de la muestra, cuando se desconoce el tamaño de la población (N), deben ser conocidos tres factores:

1. El nivel de confianza deseado, que determina el valor de z, el valor crítico de la distribución normal.
2. El error de muestreo permitido, *e*; y
3. La desviación estándar, σ; o su estimación (desviación estándar muestral)

Ejemplo 84

Una encuesta fue planeada para determinar los gastos médicos anuales de las familias de los empleados de una empresa forestal. La gerencia de la empresa desea tener 95% de confianza de que la media de la muestra esté correcta, con un margen de error de ±$50. Un estudio piloto indica que la desviación estándar es igual a $ 400.

Solución:

Recuerde que el margen de error (*e*) para la estimación por intervalo de confianza para la media, se obtiene así:

$$e = z_{(\alpha/2)} \times \frac{\sigma}{\sqrt{n}}$$, y despejando n, se obtiene la expresión: $$n = \left(\frac{z_{(\alpha/2)} \times \sigma}{e} \right)^2$$

Por lo que:

a) $$\mathrm{n} = \left(\frac{1.96 \times 400}{50} \right)^2 = 245.86, \text{ aproximadamente 246 familias.}$$

b) $$\mathrm{n} = \left(\frac{1.96 \times 400}{25} \right)^2 = 983.45, \text{ aproximadamente 984 familias.}$$

Ejemplo 85

El Instituto Nacional de Turismo de Guatemala (INGUAT) va a muestrear visitantes en las principales playas del Caribe y Pacífico, durante la Semana Santa de 2008, para estimar la proporción de extranjeros y preguntarles su opinión sobre la infraestructura existente. Las estimaciones anteriores son que el 55% de los visitantes en las playas son extranjeros.

a) ¿De qué tamaño se debe tomar la muestra para estimar la proporción de visitantes extranjeros, con precisión de 5% respecto al valor real? Use un 95% de confianza.

b) ¿Qué tamaño debe tener la muestra para que el error aumente a 10%?

Solución:

Recuerde que el margen de error (*e*) para la estimación por intervalo de confianza para la proporción, se obtiene así:

$$e = z_{(\alpha/2)} \times \sqrt{\frac{p\,(1-p)}{n}}$$, y despejando n, se tiene: $$n = \frac{z_{(\alpha/2)}^{\ 2} \times p\,(1-p)}{e^2}$$

∴ En la mayoría de las investigaciones se utiliza $e \leq 0.10$

Por lo que:

a) $$n = \frac{1.96^2 \times (0.55)\,(0.45)}{0.05^2} = 380.32,\ \text{aproximadamente 381 turistas.}$$

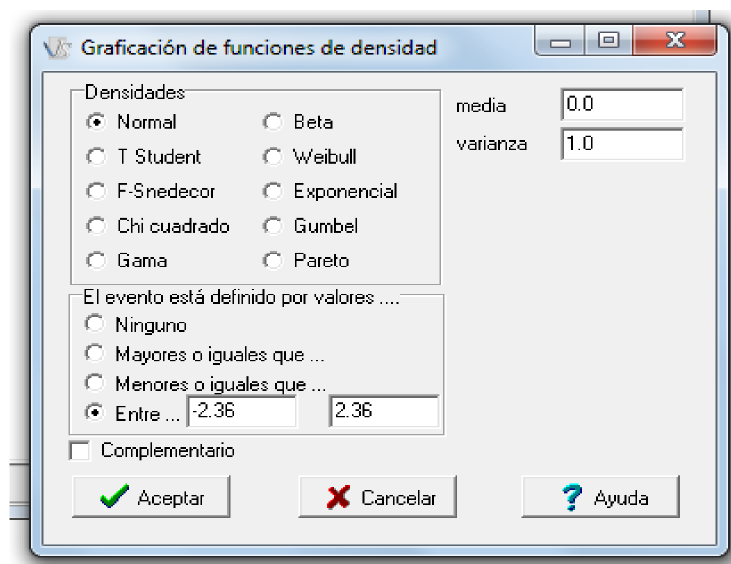
b) $$n = \frac{1.96^2 \times (0.55)\,(0.45)}{0.10^2} = 95.08,\ \text{aproximadamente 96 turistas.}$$

Ejemplo 86

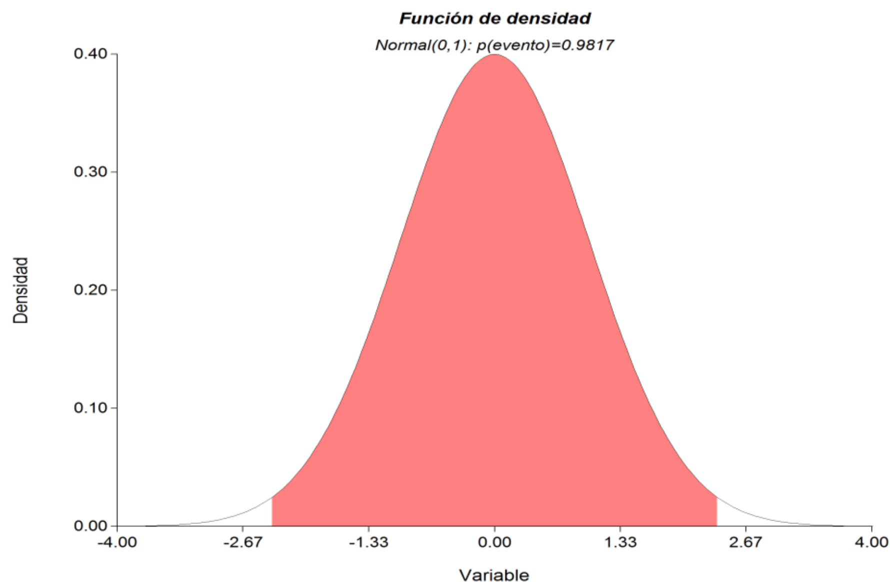
De estudios anteriores se sabe que la desviación típica de los pesos de los pollos adultos es 300 g. Queremos estimar el peso medio de los pollos adultos de una granja con un error menor que 100 g, y para ello, tomamos una muestra de 50 individuos. ¿Con qué nivel de confianza podremos realizar la estimación?

$$e = z_{(\alpha/2)} \times \frac{\sigma}{\sqrt{n}} \rightarrow 100 = z_{(\alpha/2)} \times \frac{300}{\sqrt{50}} \rightarrow z_{(\alpha/2)} = 100.\frac{\sqrt{50}}{300} \rightarrow z_{(\alpha/2)} = 2.36$$

Por lo tanto, el nivel de confianza por ser simétrico respecto a la media, se encontrará entre los valores de Z (-2.36, 2.36). En Infostat podemos calcularlo de la siguiente manera:



Que da como resultado: 0.9817 (98.17% de confianza)

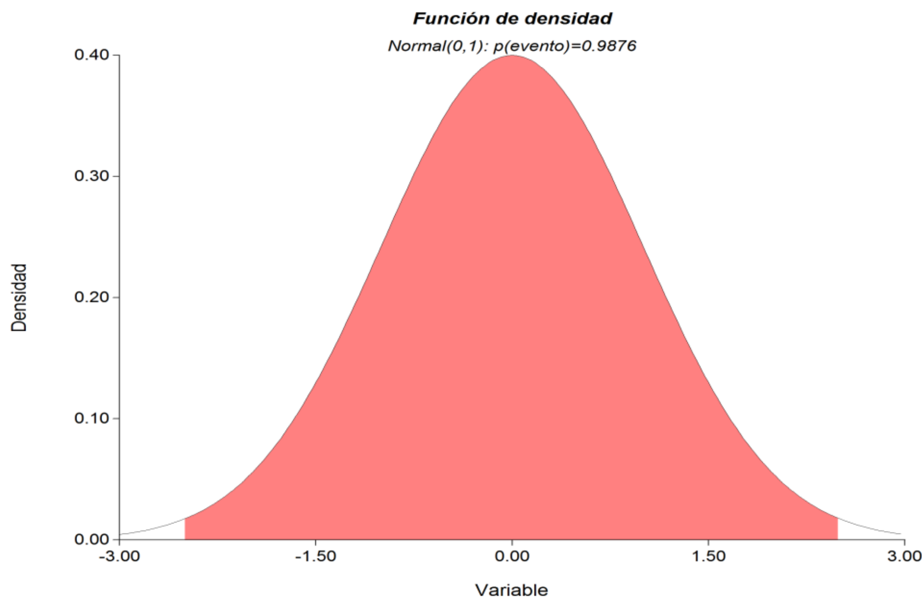


Ejemplo 87

Un agricultor quiere estimar el peso medio de las naranjas que produce, con un error menor que 10 g, utilizando una muestra de 81 naranjas. Sabiendo que la desviación típica poblacional es de 36 g, ¿cuál será el máximo nivel de confianza con que realizará la estimación?

$$e = z_{(\alpha/2)} \times \frac{\sigma}{\sqrt{n}} \rightarrow 10 = z_{(\alpha/2)} \times \frac{36}{\sqrt{81}} \rightarrow z_{(\alpha/2)} = 10.\frac{\sqrt{81}}{36} \rightarrow z_{(\alpha/2)} = 2.5$$

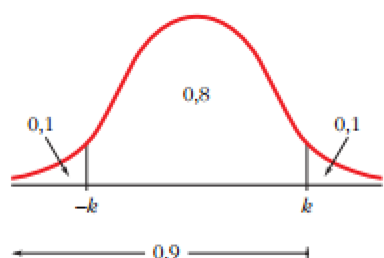
Por lo tanto, el nivel de confianza por ser simétrico respecto a la media, se encontrará entre los valores de Z (-2.5, 2.5). En Infostat nos da el siguiente resultado 0.9876 (nivel de confianza del 98.76%)



Ejemplo 88

El peso de los huevos de gallina producidos por cierta granja sigue una normal de media 65 g y desviación típica 6 g. Los huevos se clasifican en P (pequeños), M (medianos) y G (grandes). Si P supone el 10% del total y G otro 10%, ¿qué pesos marcan los límites de cada categoría?

$X \sim N(65{,}6)$. En una distribución $N(0,\ 1)$:



$$P[-k < z < k] = 0{,}8 \Rightarrow P[z < k] = 0{,}9$$

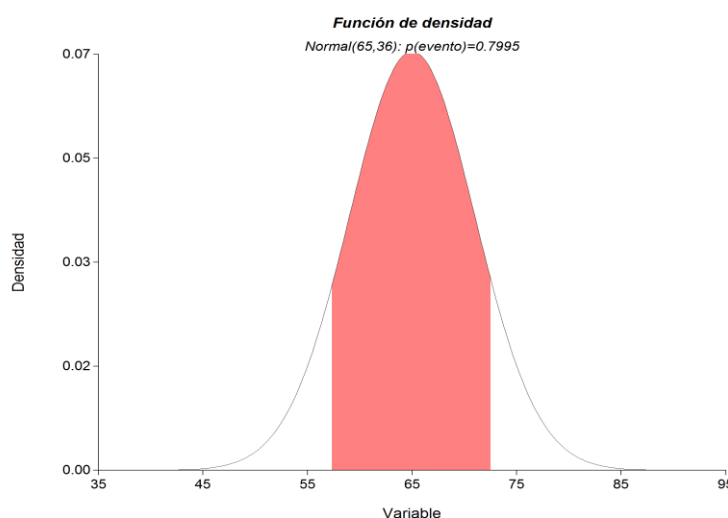
$$\Phi(k) = 0{,}9 \Rightarrow k = 1{,}28$$

$$F(k) = 0{,}9 \rightarrow \begin{cases} k = 1{,}28 \\ -k = -1{,}28 \end{cases}$$

Como z sigue una distribución $N(0,\ 1)$:

$$z = \frac{x - \mu}{\sigma} \rightarrow \begin{cases} -1{,}28 = \dfrac{x_1 - 65}{6} \rightarrow x_1 = 57{,}32 \\ 1{,}28 = \dfrac{x_2 - 65}{6} \rightarrow x_2 = 72{,}68 \end{cases}$$

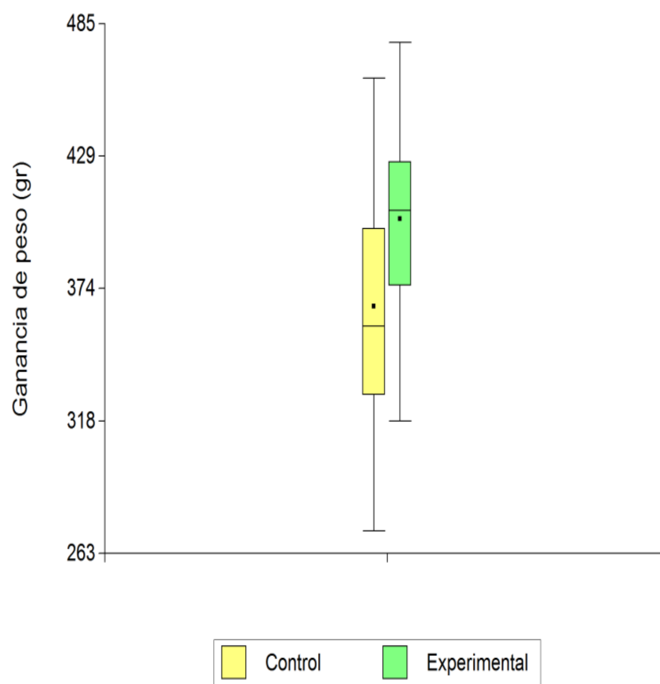
Los pesos que marcan los límites de cada categoría son 57,32 g y 72,68 g.



Ejemplo 89

El maíz común no tiene la cantidad de lisina que necesitan los animales en su pienso. Unos científicos han desarrollado ciertas variedades de maíz que contienen una mayor cantidad de lisina. En una prueba sobre la calidad del maíz con alto contenido en lisina destinada a pienso animal, un grupo experimental de 20 pollos de un día de edad empezó a recibir una ración que contenía el nuevo maíz. Un grupo control de otros 20 pollos recibió una ración que era idéntica a la anterior, con la excepción de que contenía maíz normal. Las ganancias de peso en gramos de los pollos a los 21 días fueron:

	Control	Experimental
	380	361
	283	434
	356	406
	350	427
	345	430
	321	447
	349	403
	410	318
	384	420
	455	339
	366	401
	402	393
	329	467
	316	477
	360	410
	356	375
	462	426
	399	407
	272	392
	431	326
Media	366.30	402.95
D.E.	50.81	42.73



a) Construya un intervalo de 95% de confianza para las medias de cada tipo de tratamiento. Utilice la distribución de t de Student con v=n-1 grados de libertad (gl) = 20-1= 19 gl y P(X>x) = 0.025 (α/2)

Calculador de probabilidades y cuantiles
Seleccione la distribución: Uniforme (a,b); Normal (media,varianza); T Student (v); Chi Cuadrado (v,lambda); F no central (u,v,lambda); Exponencial (lambda); Gama (lambda,r); Beta (a,b); Weibull (a,b); Doble Exponencial (a,b); Logística (a,b); Pareto (Theta,a); Gumbel (a,b); Rangos estudent.(k,v); Poisson (lambda); Binomial (n,p); Geométrica (p); Hipergeométrica (m,k,n); Binomial Negativa (m,k); Beta-Binomial(p,Rho,N)
v: 19
Valor de x: 2.0930236578
Prob. (X<=x): 0.975
Prob. (X>x): 0.025
Prob. (X=x):
Calcular | Ayuda

$$I.C.(\hat{\mu})_{1-\alpha} = \bar{x} \pm t_{(n-1,\,\alpha/2)} \times \frac{s}{\sqrt{n}}$$

$$I.C.(\hat{\mu})_{1-\alpha} = 366.3 \pm 2.093 \times \frac{50.81}{\sqrt{20}}$$

$$I.C.(\hat{\mu})_{1-\alpha} = 366.3 \pm 23.78$$

$$I.C.(\hat{\mu})_{1-\alpha} = 402.95 \pm 2.093 \times \frac{42.73}{\sqrt{20}}$$

$$I.C.(\hat{\mu})_{1-\alpha} = 402.95 \pm 20$$

b) Calcule un intervalo de 95% de confianza para las varianzas de cada tipo de tratamiento. Utilice la prueba de Ji-cuadrada con v= n-1= 20-1=19 grados de libertad, P(X>x) = 0.025 y P(X>x) = 0.975

$$\frac{(n-1)\,s^2}{\chi^2_{\left(\frac{\alpha}{2}\right)}} \le \sigma^2 \le \frac{(n-1)\,s^2}{\chi^2_{\left(1-\frac{\alpha}{2}\right)}}$$

Calculador de probabilidades y cuantiles
Seleccione la distribución
Uniforme (a,b)
Normal (media,varianza)
T Student (v)
Chi Cuadrado (v,lambda)
F no central (u,v,lambda)
Exponencial (lambda)
Gama (lambda,r)
Beta (a,b)
Weibull (a,b)
Doble Exponencial (a,b)
Logística (a,b)
Pareto (Theta,a)
Gumbel (a,b)
Rangos estudent.(k,v)
Poisson (lambda)
Binomial (n,p)
Geométrica (p)
Hipergeométrica (m,k,n)
Binomial Negativa (m,k)
Beta-Binomial(p,Rho,N)
19 v
0 lambda
Valor de x
32.8523261845
Prob. (X<=x)
0.975
Prob. (X>x)
0.025
Prob. (X=x)
Calcular
Ayuda

Calculador de probabilidades y cuantiles
Seleccione la distribución
Uniforme (a,b)
Normal (media,varianza)
T Student (v)
Chi Cuadrado (v,lambda)
F no central (u,v,lambda)
Exponencial (lambda)
Gama (lambda,r)
Beta (a,b)
Weibull (a,b)
Doble Exponencial (a,b)
Logística (a,b)
Pareto (Theta,a)
Gumbel (a,b)
Rangos estudent.(k,v)
Poisson (lambda)
Binomial (n,p)
Geométrica (p)
Hipergeométrica (m,k,n)
Binomial Negativa (m,k)
Beta-Binomial(p,Rho,N)
19 v
0 lambda
Valor de x
8.9065161347
Prob. (X<=x)
0.025
Prob. (X>x)
0.975
Prob. (X=x)
Calcular
Ayuda

$$\frac{(20-1)\,50.81^2}{32.85} \le \sigma^2 \le \frac{(20-1)\,50.81^2}{8.91} \rightarrow 1493.20 \le \sigma^2 \le 5505.22$$

$$\frac{(20-1)\,42.73^2}{32.85} \le \sigma^2 \le \frac{(20-1)\,42.73^2}{8.91} \rightarrow 1056.05 \le \sigma^2 \le 3893.51$$

LISTA DE EJERCICIOS 8

1. Una muestra aleatoria de 36 cigarrillos de una determinada marca dio un contenido promedio de nicotina de 3 miligramos. El contenido en nicotina de estos cigarrillos sigue una distribución normal con una desviación estándar de 1 miligramo.

   a) Obtenga e interprete un intervalo de confianza del 95% para el verdadero contenido promedio de nicotina en estos cigarrillos.

   b) El fabricante garantiza que el contenido promedio de nicotina es 2.9 miligramos, ¿qué puede decirse de acuerdo con el intervalo hallado?

2. Un fabricante de papel para impresoras posee un proceso de producción que opera de manera continua, a través de un turno completo de producción. Es esperado que el papel tenga un largo de 11 pulgadas, y la desviación estándar conocida sea de 0.02 pulgadas. A intervalos periódicos, son seleccionadas muestras para determinar si el largo promedio del papel aún se mantiene igual a 11 pulgadas o si algo errado ocurrió en el proceso de producción para que tenga que ser modificado el largo del papel producido. Si esta situación se presentara, se debe adoptar una acción correctiva. Una muestra aleatoria de 100 hojas fue seleccionada y se verificó que el largo promedio del papel era de 10.998 pulgadas. Con esta información obtenga una estimación por intervalo de 95% y de 99% de confianza para el largo promedio del papel en la población.

3. El gerente de control de calidad de una fábrica de bombillos de filamento necesita calcular la vida útil promedio de una gran remesa de bombillos. Se sabe que la desviación estándar del proceso es de 100 horas. Una muestra aleatoria de 50 bombillos indicó una vida útil promedio de a muestra igual a 350 horas.

   a) Construya una estimación de intervalo de 90% de confianza de la verdadera media de la vida útil de los bombillos en esta remesa.

   b) Suponga que la desviación estándar del proceso cambió a 80 horas. ¿Cuál sería su respuesta para el inciso a?

4. A partir de una población constituida por un bosque de 200 ha de *Pinus oocarpa* de 26 años, se seleccionó y midió una muestra al azar de 25 unidades de muestreo (parcelas) de 0.05 hectáreas cada una, y los volúmenes obtenidos después de procesar los datos son los siguientes:

Parcela	Volumen ($m^3.ha^{-1}$)	Parcela	Volumen ($m^3.ha^{-1}$)
1	96	14	92
2	72	15	88
3	86	16	53
4	48	17	32
5	31	18	58
6	59	19	37
7	46	20	55
8	38	21	88
9	80	22	39
10	52	23	27
11	40	24	101
12	26	25	83
13	56		

Con estos datos calcule la estimación puntual y por intervalos de 95% de confianza para el volumen promedio por hectárea.

5. Una tienda de artículos de librería recibe de su fabricante una remesa de determinada marca de plumas esferográficas. El dueño de la tienda desea calcular la proporción de plumas que presentan defectos. Una muestra aleatoria de 300 plumas es evaluada, y 30 están con defecto.

   a) Calcule una estimación con intervalo de 90% de confianza, de la proporción de plumas defectuosas en la remesa.

   b) La remesa puede ser devuelta se hubiera más del 5% con defectos; con base en los resultados de la muestra, ¿el dueño de la tienda puede devolver esa remesa?

6. Un supervisor de control de calidad en una enlatadora de frijol sabe que la cantidad exacta en cada lata varía, pues hay ciertos factores imposibles de controlar que afectan a la cantidad de llenado. El llenado medio por lata es importante, pero igualmente importante es la variación s^2 (varianza muestral) de la cantidad de llenado.

   Si s^2 es grande, algunas latas contendrán muy poco, y otras, demasiado. A fin de estimar la variación del llenado en la enlatadora, el supervisor escoge al azar 10 latas y pesa el contenido de cada una, obteniendo el siguiente pesaje (en onzas): 7.96, 7.90, 7.98, 8.01, 7.97, 8.03, 8.02, 8.04 y 8.02. Establezca un intervalo de 90% de confianza para la varianza del llenado de latas.

7. En un grupo de pacientes, el nivel de colesterol es una variable aleatoria con distribución normal, de media desconocida y varianza de 64 $(mg/ml)^2$.

   a) Para una muestra de 46 individuos que posee nivel medio de colesterol de 120mg/ml, construya el intervalo de 88% e confianza.

   b) Sí usted desea disminuir la amplitud del intervalo encontrado en el inciso anterior, ¿cuáles serían sus alternativas?

8. El intervalo [35.21; 35.99], es el intervalo de 95% de confianza, construido a partir de una muestra de tamaño 100, para la media μ de una población normal, con desviación estándar igual a 2.

   a) ¿Cuál es el valor encontrado para la media de esa muestra?

   b) Sí utilizáramos esa misma muestra, pero con una confianza del 90%, ¿Cuál sería el nuevo intervalo de confianza?

9. En la industria farmacéutica es crítica la varianza en los pesos de las medicinas. Para determinar medicina, cuyo peso se mide en gramos, una muestra de 18 unidades tuvo una varianza $s^2 = 0.36$

   a) Determine un intervalo de 95% de confianza para la varianza de la población de los pesos de esa medicina.

   b) Determine un intervalo de confianza de 90%.

10. Históricamente, 10% de un gran pedido de piezas de máquinas son defectuosas. Si fueran seleccionadas muestras aleatorias de 400 piezas, qué proporción de las muestras tendrá:

a) Entre 9% y 12% de piezas defectuosas?
b) Menos de 8% de piezas defectuosas?
c) Si un tamaño solamente de 100 piezas fuera seleccionado, cuáles serán las respuestas para los incisos a y b?

11. En el *1997 Statistical Abstract of the United States* se menciona el porcentaje de personas de 18 años o más que fuman. Suponga que un estudio se diseña para reunir nuevos datos de fumadores y no fumadores. El mejor estimado preliminar de la proporción poblacional de quienes fuman es de 30%.

    a) ¿De qué tamaño debe tomarse una muestra para estimar la proporción de fumadores en la población, considerando un margen de error igual al 2%? Utilice un nivel de 95% de confianza.

    b) Suponga que el estudio usa su recomendación de tamaño de muestra del inciso a, y ve que hay 520 fumadores. ¿Cuál es la estimación puntual de la proporción de fumadores en la población?

    c) ¿Cuál es el intervalo de 95% de confianza para la proporción de fumadores en la población? Utilice los datos del inciso b.

12. Un investigador de mercados de una empresa grande de electrodomésticos realizará un estudio sobre los hábitos relativos a ver televisión de los adultos de la ciudad de Huehuetenango. Una muestra aleatoria de 40 entrevistados es seleccionada, brindando los siguientes resultados relacionados con el tiempo que utiliza para ver televisión:
    $\bar{x}$ = 15.3 horas por semana y s = 3.8 horas. Con esta información:

    a) Construya un intervalo de 95% de confianza para la cantidad media de tiempo que se utiliza para ver televisión, por semana, en esa ciudad.

    b) Si el investigador de mercados desea realizar otra encuesta, en una ciudad cercana, ¿qué tamaño de muestra será necesario si él desea tener 95% de confianza de tener un margen de error de ± 2 horas a partir del supuesto que la desviación estándar de esa población es de 5 horas?

13. Se desea realizar una encuesta entre la población juvenil de una determinada localidad para determinar la proporción de jóvenes que estaría a favor de una nueva zona de ocio. El número de jóvenes de dicha población es N=2,000. Determinar el tamaño de muestra necesario para estimar la proporción de estudiantes que están a favor con un error de estimación de 0.05 y un nivel de confianza del 95%.

14. Una máquina llena cajas con cierto cereal. El supervisor desea conocer con un error de estimación de máximo 0.1 y un nivel de confianza del 90%, una media estimada del peso. Como la varianza era desconocida se procedió a escoger una muestra piloto. Los resultados fueron los siguientes (expresados en onzas): 11.02, 11.14, 10.78, 11.59, 11.58, 11.19, 11.71, 11.27, 10.93, 10.94. ¿Cuántas cajas debe escoger para que se cumplan los requisitos propuestos?

15. Un estudiante de EPSA realizó una encuesta en una comunidad y entrevistó a 150 familias de una determinada población, encontró que en 25 de ellas había tres o más hijos. Calcule el intervalo de confianza para estimar la proporción real de las familias en las que hay tres o más hijos, con un nivel de confianza del 90%.

16. En ciudad se sabe que el porcentaje de habitantes con estudios universitarios se sitúa alrededor de 15%. Se desea actualizar los datos y, para ello, se va a extraer una muestra aleatoria de tamaño *n* para hacer la estimación del porcentaje actual. ¿De qué tamaño mínimo deberemos seleccionar la muestra para que le error en la estimación de la proporción sea menor de 0.02, con un nivel de confianza del 95.44%?

17. De 1500 personas encuestadas en un sondeo preelectoral, 800 manifiestan su intención de votar. ¿Entre qué valores puede estimarse, con un 95% de confianza, que se encontrará el nivel de abstención en el conjunto del censo?

18. Para estimar la proporción de las familias de una determinada ciudad que poseen microondas, se va a tomar una muestra aleatoria de tamaño *n*. Calcule el mínimo valor de *n* para garantizar que, con un nivel de 95% de confianza, el error en la estimación sea menor que 0.05. (Ya que se desconoce la proporción, se tiene que tomar el caso más desfavorable de que sea 0.5).

19. Por medio de una encuesta realizada a 800 personas sobre la elección de alcalde de una ciudad, se estimó que la proporción de votantes al candidato A estaba entre el 54% y el 59%. ¿Con qué nivel de confianza se realizó la estimación?

20. En cierto instituto de Enseñanza Secundaria hay matriculados 800 alumnos. A una muestra seleccionada aleatoriamente de un 15% de ellos, se les preguntó si utilizaban la cafetería del instituto. Contestaron negativamente un total de 24 alumnos. Calcule el intervalo de confianza del 99% para estimar la proporción de alumnos que utilizan la cafetería del instituto.

21. Un investigador está interesado en estimar la ganancia en peso total, en 0 a 4 semanas de 1,000 pollitos alimentados con una ración. Obviamente, pesar cada vez sería tedioso y llevaría demasiado tiempo. Por lo tanto se debe determinar el número de pollitos a seleccionar en una muestra, para estimar el total con un límite para el error de estimación igual a 1,000 gramos. Muchos estudios similares sobre nutrición de pollitos se han llevado a cabo en el pasado. Usando datos de estos estudios, el investigador encontró que la varianza es aproximadamente de 36 gramos .Determinar el tamaño de muestra requerido con confianza del 95%

22. Se pesaron 35 pollos de granja y sus pesos , en libras, fueron los siguientes:

    4.3 – 5.2 – 4.3 – 4.7 –3.9 – 4.2 – 5 – 4.1 –3.9 – 4.4 – 4.7 – 3.7 –3.5 – 5.1 –4.3 – 4.6 –3.6 – 5.5 – 4.2– 4.7 3.9 – 5.2 – 4.3 – 4.7 –3.9 – 5.2 – 4.3 – 4.7 – 3.9 – 5.2 – 4.3 – 4.7 –3.9 – 5.2 – 4.3

    Construya un intervalo de confianza del 97% para el verdadero peso medio de los pollos de la granja.

23. Tomada, al azar, una muestra de 120 estudiantes de una Universidad, se encontró que 54 de ellos hablaban inglés. Halle, con un nivel de confianza del 90%, un intervalo de confianza para estimar la proporción de estudiantes que hablan el idioma inglés entre los estudiantes de esa Universidad.

UNIDAD V
PRUEBAS DE HIPOTESIS

5.1 DEFINICIONES BASICAS

5.1.1 HIPOTESIS

Una hipótesis estadística es una suposición o afirmación sobre los parámetros de una o más poblaciones. La veracidad o falsedad de una hipótesis estadística *nunca* es conocida con certeza, a menos que, se examine a toda la población, lo que es imprráctico en la mayor parte de las situaciones.

De esta forma, se toma una muestra aleatoria de la población de interés y con base en esta muestra es establecido sí la hipótesis es probablemente verdadera o probablemente falsa. La decisión de que la hipótesis es probablemente verdadera o falsa es tomada con base en distribuciones de probabilidad denominadas: "**distribuciones muestrales**". En Estadística se trabaja con dos tipos de hipótesis:

a) La **hipótesis nula,** es la hipótesis de igualdad. Esta hipótesis es denominada hipótesis de nulidad y es representada por H_0. La hipótesis nula es normalmente formulada con el objetivo de ser rechazada. El rechazo de la hipótesis nula conduce a la aceptación de otra hipótesis denominada: alternativa o alterna.

b) La **hipótesis alterna** es la definición operacional de la hipótesis de la investigación que se desea comprobar. La naturaleza del estudio irá a definir como debe ser formulada la hipótesis alternativa. Por ejemplo, sí la prueba es del tipo paramétrico, donde el parámetro a ser evaluado es representado por θ, entonces la hipótesis nula sería: $H_0 : \theta = \theta_0$ y las hipótesis alternativas serian:

H_a: $\theta \neq \theta_0$,
H_a: $\theta > \theta_0$, o
H_a: $\theta < \theta_0$.

En el primer caso, H_a: $\theta \neq \theta_0$, se dice que es una prueba bilateral (bicaudal o de dos colas), por otra parte, sí H_a: $\theta > \theta_0$, se dice que la prueba es unilateral (de una cola o unicaudal) a la derecha, y sí H_a: $\theta < \theta_0$, entonces, se dice que la prueba es unilateral (de una cola o unicaudal) a la izquierda.

5.1.2 TIPOS DE ERRORES EN LA CONCLUSION DE UNA PRUEBA DE HIPOTESIS

Los dos errores que pueden ser cometidos cuando se realiza una prueba de hipótesis son:

a) Rechazar la hipótesis nula, cuando tal hipótesis realmente es verdadera, y

b) Aceptar la hipótesis nula, cuando ella en realidad es falsa.

Note que ningún error es cometido y la conclusión es correcta, cuando se rechaza la hipótesis nula y ella es realmente falsa, o cuando decidimos aceptarla en el caso de que ella sea realmente verdadera. Al error citado en el inciso a) se le denomina: **error tipo I**, en tanto que al error citado en el inciso b) se le conoce como: **error tipo II**. La siguiente figura resume estas posibles situaciones:

		Situación en la población	
		H_0 verdadera	H_0 falsa
Conclusión de la prueba	Aceptar H_0	Decisión correcta	Error Tipo II
	Rechazar H_0	Error Tipo I	Decisión correcta

Una parte importante de la prueba de hipótesis, se refiere al control de la probabilidad de cometer el error tipo I. Esa probabilidad es denotada por la letra griega α. Por otra parte, la probabilidad de cometer el error tipo II, se designa con la letra griega β. Esto es,

α = P (error tipo I) = P (rechazar $H_0 \mid H_0$ es verdadera);
β = P (error tipo II) = P (aceptar $H_0 \mid H_0$ es falsa).

5.1.3 NIVEL DE SIGNIFICANCIA

El nivel de significancia de una prueba es la probabilidad máxima que estamos dispuestos a aceptar, de cometer un error tipo I.

5.1.4 NIVEL DE CONFIANZA

El complemento ($1-\alpha$) de la probabilidad de un error tipo I es llamado coeficiente de confianza, que, al ser multiplicado por 100, produce el nivel de confianza.

El coeficiente de confianza es definido como la probabilidad de que la hipótesis nula no sea rechazada cuando de hecho sea verdadera y no debe ser rechazada. En términos de la metodología de la prueba de hipótesis, ese coeficiente representa la probabilidad de que se concluya que el determinado valor del parámetro que está siendo evaluado con la prueba de hipótesis sea admisible.

5.2 PASOS PARA LA EVALUACION DE UNA HIPOTESIS ESTADISTICA

A continuación se presenta un resumen de los pasos que se pueden aplicar en cualquier prueba de hipótesis.

1. Defina las hipótesis nula y alternativa adecuadas para el caso.
   Por ejemplo, si estamos interesados en una prueba de hipótesis referente a los valores de una media μ de población debe asumir una de las tres formas siguientes:

Ho: $\mu \geq \mu_o$ Ha: $\mu < \mu_o$	Ho: $\mu \leq \mu_o$ Ha: $\mu > \mu_o$	Ho: $\mu = \mu_o$ Ha: $\mu \neq \mu_o$
(a)	**(b)**	**(c)**

En los casos **a** y **b** se dice que la prueba es unilateral (de una cola o unicaudal) y en el caso **c** se trata de una prueba bilateral (bicaudal o de dos colas).

2. Seleccione el estadístico de prueba que utilizará para decidir si rechazar o no la hipótesis nula. El estadístico de prueba es un número que se utiliza para la toma de decisiones en la prueba de hipótesis. Por ejemplo:

Prueba de hipótesis	**Estadístico de prueba**
Acerca de una media poblacional	t de Student o z (distribución normal)
Acerca de la diferencia entre dos medias independientes	t de Student
Acerca de la diferencia entre dos medias dependientes (pareadas)	t de Student
Acerca de la varianza de una población normal	χ^2 (Ji-cuadrada)
Para la comparación entre dos varianzas	F de Fisher–Snedecor
Acerca de una proporción poblacional	z (distribución normal)
Para la comparación entre dos proporciones poblacionales	z (distribución normal)

3. Especifique el nivel de significancia α, para la prueba.

   En la práctica, la persona que efectúa la prueba de hipótesis especifica la máxima probabilidad permisible, llamada nivel de significancia para la prueba, de cometer un error de tipo I. Los valores más usuales de α son: 0.01, 0.05 y 0.10.

4. Use el nivel de significancia α para establecer la regla de rechazo que indique los valores del estadístico de prueba que conducirán al rechazo de H_0.

5. Reúna los datos de la muestra aleatoria y representativa y calcule el valor del estadístico de prueba.

6. a. Compare el valor del estadístico de prueba con el o los valores críticos especificados en la regla de rechazo, para determinar si H_0 se debe rechazar o no. O bien.

   b. Calcule el valor p basado en el estadístico de prueba en el paso 5. Use el valor p para determinar si se debe rechazar o no H_0. Como regla general, se rechaza H_0 si $p < \alpha$, en todas las pruebas de hipótesis.

7. Conclusiones, en términos prácticos, en función de lo evaluado.

5.3 PRUEBA DE HIPÓTESIS ESTADÍSTICA ACERCA DE UNA MEDIA POBLACIONAL NORMAL, CON VARIANZA (σ^2) DESCONOCIDA. MUESTRAS PEQUEÑAS (n < 30)

Ejemplo 90

El fabricante del cereal "*Coco flakes*" afirma que el peso promedio de cada caja de una presentación del cereal es de 500 gramos. Para ello tomó el peso a una muestra aleatoria y representativa de 16 cajas. Pruebe con un nivel de significancia α = 0.05, si la afirmación del fabricante es verdadera.

506, 508, 499, 503, 504, 510, 497, 512,
514, 505, 493, 496, 506, 502, 509, 496.

Solución:

Datos: n = 16, $\bar{x}$ = 503.75 gramos, s = 6.20 gramos, α = 0.05

1. Establecer las hipótesis

H_0 : μ = 500 gramos (el peso promedio de todas las cajas de cereal es igual a 500 gr.)
H_a : $\mu \neq$ 500 gramos (el peso promedio de todas las cajas de cereal es diferente a 500 gr.)

2. Cálculo de la estadística de prueba.

Como la muestra es pequeña, se utiliza la estadística t de Student.

$$t_o = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \qquad t_o = \frac{503.75 - 500}{6.20/\sqrt{16}} = \frac{3.75}{1.55} = 2.42$$

3. Definir las regiones de rechazo y de aceptación de Ho.

El valor crítico de la estadística va a definir el límite entre las regiones de rechazo (o no aceptación) y la de aceptación (o de no rechazo) de la hipótesis nula. Este valor se consulta en la tabla t de Student, considerando n – 1 grados de libertad y un valor α = 0.05, para una prueba bilateral (o de dos colas), ya que la hipótesis alterna fue definida en términos de desigualdad.

Para este caso, tenemos 16– 1 = 15 grados de libertad, y al consultar la tabla de t de Student, obtenemos un valor de t crítico = 2.131.

			Tabla t de Student			
	0.10	0.05	0.025	0.01	0.005	**Una cola**
gl	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**Dos colas**
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
4	1.533	2.132	2.776	3.747	4.604	
5	1.476	2.015	2.571	3.365	4.032	
6	1.440	1.943	2.447	3.143	3.707	
7	1.415	1.895	2.365	2.998	3.499	
8	1.397	1.860	2.306	2.896	3.355	
9	1.383	1.833	2.262	2.821	3.250	
10	1.372	1.812	2.228	2.764	3.169	
11	1.363	1.796	2.201	2.718	3.106	
12	1.356	1.782	2.179	2.681	3.055	
13	1.350	1.771	2.160	2.650	3.012	
14	1.345	1.761	2.145	2.624	2.977	
15	1.341	1.753	2.131	2.602	2.947	
16	1.337	1.746	2.120	2.583	2.921	
17	1.333	1.740	2.110	2.567	2.898	
18	1.330	1.734	2.101	2.552	2.878	
19	1.328	1.729	2.093	2.539	2.861	
20	1.325	1.725	2.086	2.528	2.845	

Gráficamente se presenta a continuación las regiones de rechazo y de aceptación de Ho.



Región de rechazo de Ho

Región de aceptación de Ho

–2.131 2.131

4. Regla de decisión

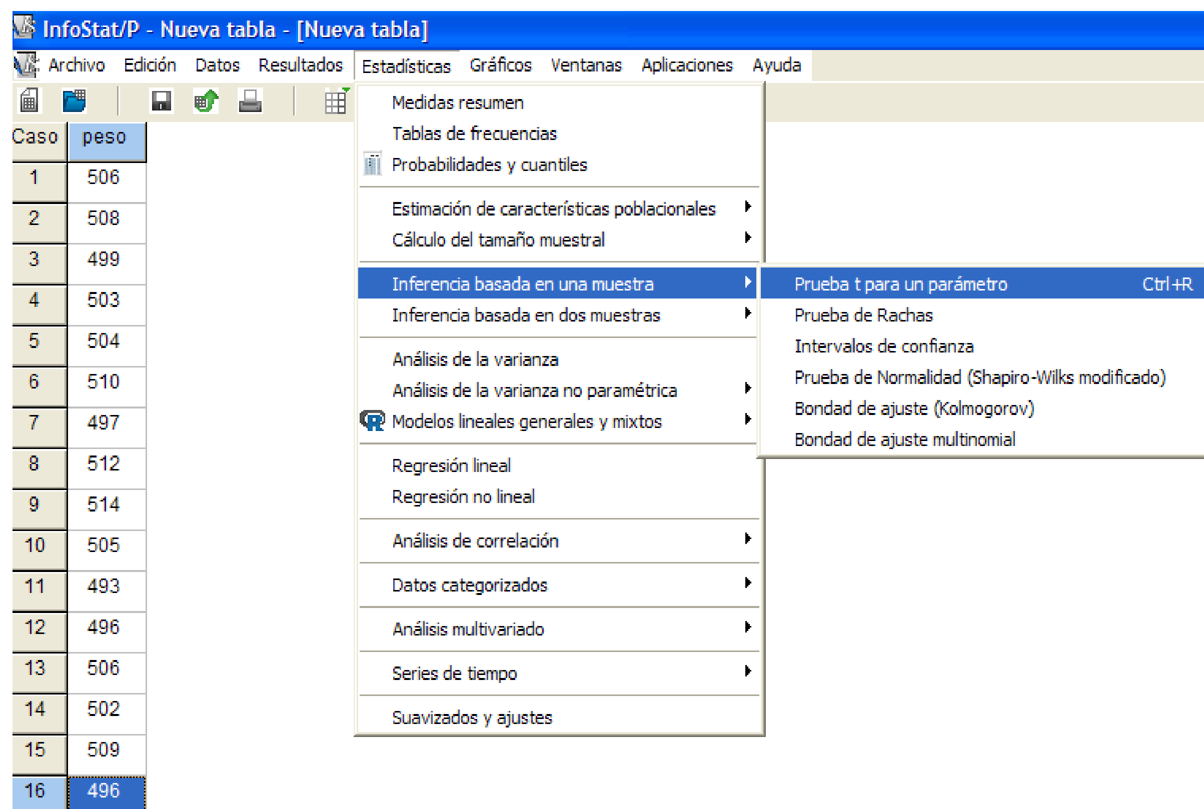
Sí el valor absoluto de t_o, calculado en el paso 2, es mayor que el valor de t crítico se rechaza la H_o, en caso contrario, H_0 se acepta.

5. Conclusión

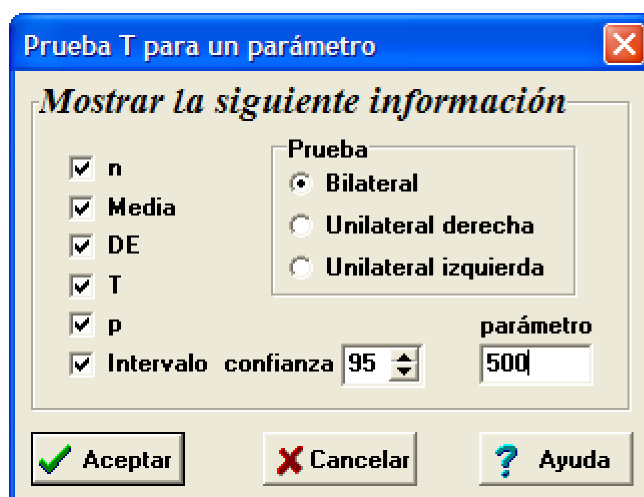
Como el valor de t_o = 2.42 es mayor que el valor de t crítico (2.131), se rechaza la hipótesis nula y concluímos que estadísticamente el peso promedio de las cajas de cereal “*Coco flakes*” es diferente a 500 gramos.

Resolución del ejemplo 90 con Infostat

1. Ingreso de datos



2. Solicitar análisis



3. Verificar resultados y comparar con los cálculos manuales

Prueba T para un parámetro

Valor del parámetro probado: 500

Variable	n	Media	DE	LI(95)	LS(95)	T	p(Bilateral)
peso	16	503.75	6.20	500.45	507.05	2.42	0.0288

Ejemplo 91

El tiempo medio, por operario, para ejecutar una tarea, ha sido de 100 minutos. Se introdujo un nuevo método para disminuir este tiempo, y, luego de cierto período, se sorteó una muestra de 16 operarios, midiendo el tiempo de ejecución empleado por cada uno. El tiempo medio de la muestra fue de 85 minutos con desviación estándar de 12 minutos. ¿Considera que este resultado evidencia una mejora en el tiempo empleado para realizar la tarea? Presente las conclusiones utilizando niveles de significancia de 5% y 1% e indique cuáles son las suposiciones teóricas necesarias que deben ser hechas para resolver el problema.

Solución.

Datos: n = 16, $\bar{x}$ = 85 minutos, s = 12 minutos, α = 0.05 y 0.01

1. Establecer las hipótesis

H_o : $\mu \geq 100$ minutos (el tiempo medio, por operario, para ejecutar una tarea, es mayor o igual a 100 minutos)

H_a : $\mu < 100$ minutos (el tiempo medio, por operario, para ejecutar una tarea ha mejorado, respecto al tiempo actual 100 minutos, al utilizar el nuevo método)

2. Cálculo de la estadística de prueba.

Como la muestra es pequeña, se utiliza la estadística t de Student.

$$t_o = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \qquad \frac{85 - 100}{12/4} = -5$$

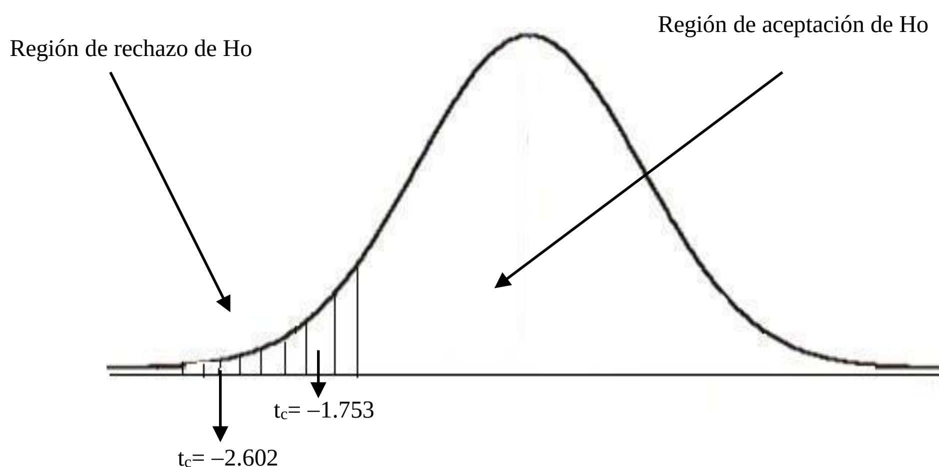
3. Definir las regiones de rechazo y aceptación de Ho.

Tabla t de Student

	0.10	0.05	0.025	0.01	0.005	Una cola
gl	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**Dos colas**
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
4	1.533	2.132	2.776	3.747	4.604	
5	1.476	2.015	2.571	3.365	4.032	
6	1.440	1.943	2.447	3.143	3.707	
7	1.415	1.895	2.365	2.998	3.499	
8	1.397	1.860	2.306	2.896	3.355	
9	1.383	1.833	2.262	2.821	3.250	
10	1.372	1.812	2.228	2.764	3.169	
11	1.363	1.796	2.201	2.718	3.106	
12	1.356	1.782	2.179	2.681	3.055	
13	1.350	1.771	2.160	2.650	3.012	
14	1.345	1.761	2.145	2.624	2.977	
15	1.341	1.753	2.131	2.602	2.947	

Para este caso, tenemos 16– 1 = 15 grados de libertad, y al consultar la tabla de t de Student, obtenemos un valor de t crítico = −1.753, cuando consideramos un valor de α = 5% , y −2.602 cuando consideramos un valor de α = 1%.

Gráficamente se presenta a continuación las regiones de rechazo y de aceptación de Ho.



4. Regla de decisión

Sí el valor de $-t_o$ es menor que el valor de t crítico, se rechaza la H_o, en caso contrario, se acepta H_0.

5. Conclusión

Como el valor de t observado (−5) se encuentra dentro de la región de rechazo de la hipótesis nula, con los niveles de 5% y 1% de significancia, se puede afirmar que el tiempo medio, por operario, para ejecutar una tarea ha mejorado, respecto al tiempo actual 100 minutos, al utilizar el nuevo método.

5.4 PRUEBA DE HIPÓTESIS ESTADÍSTICA ACERCA DE UNA MEDIA POBLACIONAL NORMAL, CON VARIANZA (σ^2) CONOCIDA.

Ejemplo 92

La Asociación de Propietarios de Industrias Lácteas de Guatemala (ASOLAC) está preocupada con el tiempo perdido en accidentes de trabajo, cuya media, en los últimos 5 años, ha sido de 60 horas/hombre/año con desviación estándar (σ) de 20 horas/hombre. Se implementó un programa de prevención de accidentes con apoyo del Instituto Guatemalteco de Seguridad Social (IGSS) y, luego de su aplicación, se tomó una muestra de 9 industrias y se midió el número de horas/hombre/año perdidas por accidente, que fue de 50 horas.

¿Podrá afirmarse, con un nivel de significancia del 5%, que existe evidencia de mejoría en cuanto a la reducción del número de horas perdidas por causa de los accidentes?

Solución.

Datos: $n = 9$ industrias, $\bar{x} = 50$ horas, $\sigma = 20$ horas, $\alpha = 0.05$

1. Establecer las hipótesis

H_0: $\mu \geq 60$ (el tiempo promedio perdido en accidentes de trabajo es mayor o igual a 60 horas/hombre/año)

H_a: $\mu < 60$ (el tiempo promedio perdido en accidentes de trabajo es menor a 60 horas/hombre/año)

2. Cálculo de la estadística de prueba.

A pesar de que la muestra es pequeña, se utiliza la estadística z de la distribución normal, porque se conoce el valor de la varianza poblacional. Para esto es necesario estandarizar el resultado muestral, de la siguiente manera:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{(50 - 60)}{20 / \sqrt{9}} = -1.50$$

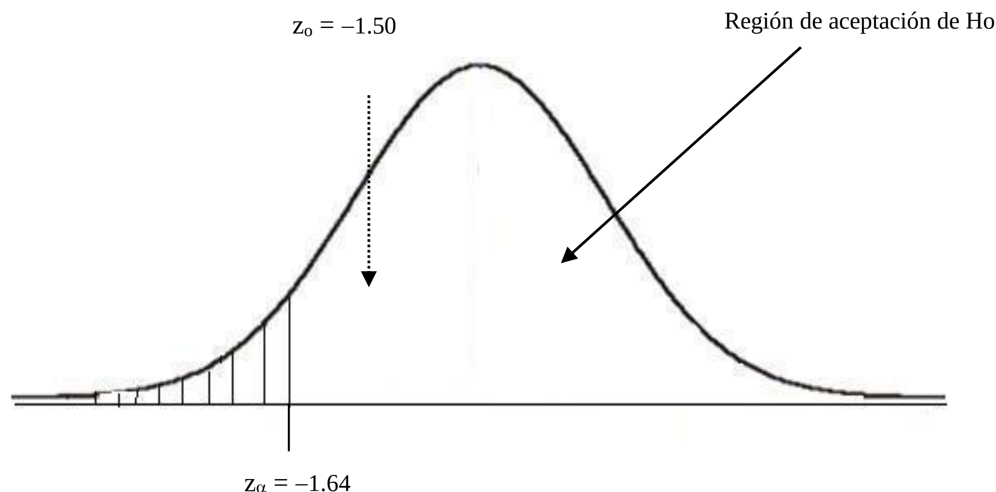
3. Definir las regiones de rechazo y aceptación de Ho.

Para saber si el valor de la estadística de z observado o calculado (−1.50) es poco probable, es necesario compararlo con el valor crítico $-z_\alpha$ (pues se trata de una prueba unilateral a la izquierda). En el siguiente cuadro se presentan algunos valores críticos de z dependiendo del nivel de significancia con el que se esté trabajando:

Valores de z para algunos niveles de significancia

	α = Nivel de significancia		
	10%	**5%**	**1%**
Prueba bilateral	1.64	1.96	2.57
Prueba unilateral	1.28	1.64	2.33

Para este caso, z_α equivale a −1.64, ya que el nivel de significancia fue fijado en 5%. Gráficamente se presenta a continuación las regiones de rechazo y de aceptación de Ho.



4. Regla de decisión

Se rechaza la hipótesis nula H_0 sí el valor de z observado es:

(i) mayor que z_α (en la prueba unilateral a la derecha);
(ii) menor que $-z_\alpha$ (en la prueba unilateral a la izquierda) y
(iii) mayor que $z_{\alpha/2}$ o menor que $-z_{\alpha/2}$ (en la prueba bilateral).

5. Conclusión

Debido a que el valor de z observado se ubica en la región de aceptación de la hipótesis nula, se concluye que no es posible afirmar, con un nivel de significancia del 5%, que el programa de prevención de accidentes haya dado resultado.

5.5 PRUEBA DE HIPÓTESIS ESTADÍSTICA ACERCA DE UNA PROPORCIÓN POBLACIONAL

Ejemplo 93

El restaurante de comida rápida "Burger Queen" planea una oferta especial que permita a los clientes comprar vasos de diseño especial con personajes de la película infantil "Los Increíbles". Si más del 15% de los clientes compran esos vasos, se implantará la promoción. En una prueba preliminar en varios locales de la ciudad capital de Guatemala, 88 de 500 clientes los compraron. ¿Se debe implantar la promoción especial? Lleve a cabo una prueba de hipótesis que apoye su decisión. Use un nivel de significancia de 0.05. ¿Cuál es su recomendación?

Solución.
Datos: $n = 500$, $p_o = 0.15$ $\bar{p} = 88/500 = 0.18$

1. Establecer las hipótesis

H_0: $p_o \leq 0.15$ (15% o menos de los clientes compran los vasos promocionales)
H_a: $p_o > 0.15$ (más del 15% de los clientes compran los vasos promocionales)

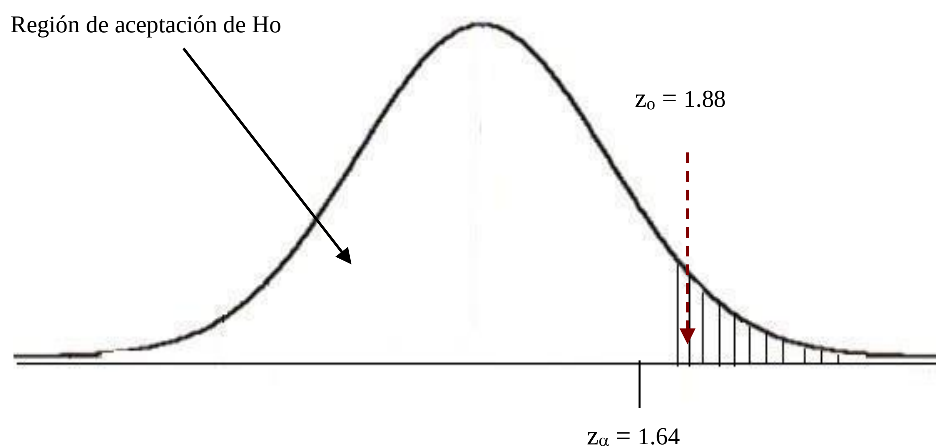
2. Cálculo de la estadística de prueba.

En este caso se utiliza la estadística z de la distribución normal:

$$z = \frac{\bar{p} - p_o}{\sigma_{\bar{p}}} = \frac{\bar{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} = \frac{0.18 - 0.15}{\sqrt{\frac{0.15(1-0.15)}{500}}} = \frac{0.03}{0.016} = 1.88$$

3. Definir las regiones de rechazo y aceptación de Ho.

Para saber si el valor de la estadística de z observado o calculado (1.88) es poco probable, es necesario compararlo con el valor crítico z_α (pues se trata de una prueba unilateral a la derecha). Para este caso, z_α equivale a 1.64, ya que el nivel de significancia fue fijado en 5%. En la gráfica siguiente se ilustran estos resultados.



4. Regla de decisión

Se rechaza la hipótesis nula H_0 sí el valor de z_o es mayor o igual que el z crítico.

5. Conclusión

Debido a que el valor de z observado se ubica en la región de rechazo de la hipótesis nula, se concluye, con un nivel de significancia del 5%, que estadísticamente más del 15% de los clientes compran los vasos promocionales, por lo que el restaurante debe implantar la promoción especial.

Ejemplo 94

El gerente de la cadena de restaurantes de comida rápida "Pollo Tejano" afirma que solamente el 30% de los clientes individuales consume el menú de 3 piezas. Suponga que una muestra de 480 clientes en restaurantes indicó que 128 solicitan ese tipo de menú. Pruebe la afirmación del gerente a un nivel de significancia de 0.05. ¿Cuál es su conclusión?

Solución.

Datos: $n = 480$, $p_o = 0.30$ $\bar{p} = 128/480 = 0.27$

1. Establecer las hipótesis

H_0: $p_o = 0.30$ (el 30% de los clientes individuales consume el menú de 3 piezas)
H_a: $p_o \neq 0.30$ (una cantidad de clientes diferente a 30% consume el menú de 3 piezas)

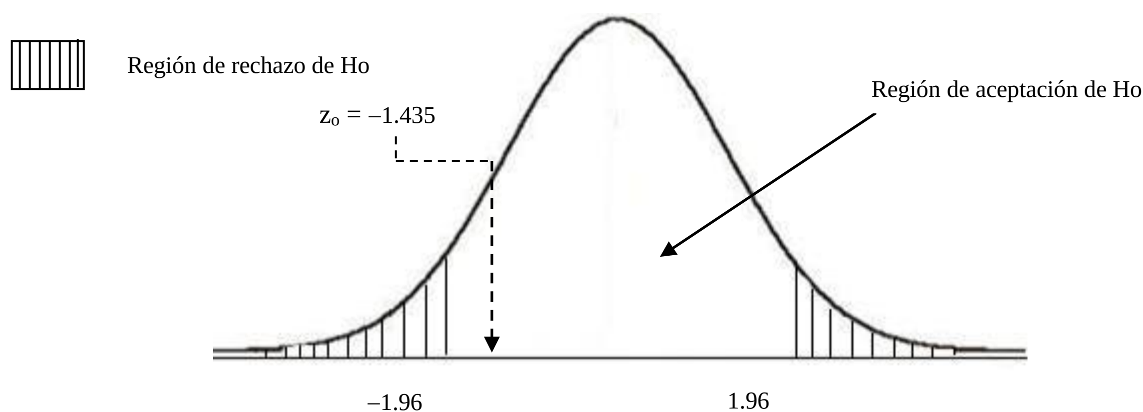
2. Cálculo de la estadística de prueba.

En este caso se utiliza la estadística z de la distribución normal:

$$z = \frac{\bar{p} - p_o}{\sigma_{\bar{p}}} = \frac{\bar{p} - p_o}{\sqrt{\frac{p_o(1 - p_o)}{n}}} = \frac{0.27 - 0.30}{\sqrt{\frac{0.30(1 - 0.30)}{480}}} = \frac{-0.03}{0.0209} = -1.435$$

3. Definir las regiones de rechazo y aceptación de Ho.

Para este caso, el valor crítico de $z_{\alpha/2}$ equivale a 1.96, por tratarse de una prueba bilateral. Gráficamente se presenta a continuación las regiones de rechazo y de aceptación de Ho.



4. Regla de decisión

Se acepta la hipótesis nula H_0 sí el valor de z_o está comprendido entre el rango [–1.96; 1.96], en caso contrario, se rechaza.

5. Conclusión

Debido a que el valor de z observado se ubica en la región de aceptación de la hipótesis nula, se concluye, con un nivel de significancia del 5%, que únicamente el 30% de los clientes individuales consume el menú de 3 piezas de pollo.

5.6 PRUEBA DE HIPÓTESIS ESTADÍSTICA ACERCA DE UNA VARIANZA POBLACIONAL

Ejemplo 95

Una de las maneras de verificar la calidad de un producto es controlando su variabilidad. Una máquina para empacar café soluble, se considera que está bien regulada para llenar los paquetes, si tiene una desviación estándar de 10 g y media de 500g y donde el peso de cada paquete se distribuye normalmente. Colectada una muestra de n = 16, se observó una varianza de 169 g^2. ¿Es posible afirmar con este resultado que la máquina no está regulada en función de la variabilidad, suponiendo un nivel de significancia del 5%?

Solución

Datos:

n = 16, $\bar{x}$ = 500 gramos, σ^2 = 169 gramos^2, α = 0.05

1. Establecer las hipótesis

H_0: $\sigma^2 = 100$ (la máquina para empacar café tiene una varianza igual a 100 g^2, por lo que está bien regulada)

H_a: $\sigma^2 \neq 100$ (la máquina para empacar café tiene una varianza distinta a 100 g^2)

2. Cálculo de la estadística de prueba.

En este caso se utiliza la estadística χ^2 (Ji cuadrada):

$$\chi^2_{o(n-1)} = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(16-1)169}{100} = 25.35$$

Quiere decir que el cociente arriba descrito tiene una distribución Ji–cuadrada con "n–1" grados de libertad. La Ji–cuadrada es una distribución asimétrica positiva que varía de cero a más infinito.

3. Definir las regiones de rechazo y aceptación de Ho.

En función del tipo de hipótesis alternativa se define la región de rechazo. Se pueden tener las siguientes opciones de hipótesis:

Ho: $\sigma^2 \geq \sigma^2_o$ Ha: $\sigma^2 < \sigma^2_o$	Ho: $\sigma^2 \leq \sigma^2_o$ Ha: $\sigma^2 > \sigma^2_o$	Ho: $\sigma^2 = \sigma^2_o$ Ha: $\sigma^2 \neq \sigma^2_o$

4. Regla de decisión.

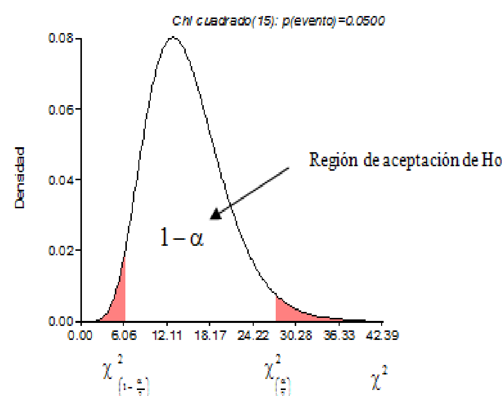
En la siguiente figura se ilustra la zona de rechazo y la de aceptación de la hipótesis nula:

Se rechaza la hipótesis nula H_0 sí el valor de χ^2 observado es:

a) mayor que $\chi^2_{\left(\frac{\alpha}{2}\right)}$ (en la prueba unilateral a la derecha);

b) menor que $\chi^2_{\left(1-\frac{\alpha}{2}\right)}$ (en la prueba unilateral a la izquierda)

c) mayor que $\chi^2_{\left(\frac{\alpha}{2}\right)}$ o menor que $\chi^2_{\left(1-\frac{\alpha}{2}\right)}$ (en la prueba bilateral).



Este valor tabular se obtuvo así:

			Área correspondiente al extremo derecho de una distribución Jí-cuadrada.					
gl	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025
1	0.00004	0.00016	0.00098	0.00393	0.01579	2.70554	3.84146	5.02389
2	0.01003	0.02010	0.05064	0.10259	0.21072	4.60517	5.99146	7.37776
3	0.07172	0.11483	0.21580	0.35185	0.58437	6.25139	7.81473	9.34840
4	0.20699	0.29711	0.48442	0.71072	1.06362	7.77944	9.48773	11.14329
5	0.41174	0.55430	0.83121	1.14548	1.61031	9.23636	11.07050	12.83250
6	0.67573	0.87209	1.23734	1.63538	2.20413	10.64464	12.59159	14.44938
7	0.98926	1.23904	1.68987	2.16735	2.83311	12.01704	14.06714	16.01276
8	1.34441	1.64650	2.17973	2.73264	3.48954	13.36157	15.50731	17.53455
9	1.73493	2.08790	2.70039	3.32511	4.16816	14.68366	16.91898	19.02277
10	2.15586	2.55821	3.24697	3.94030	4.86518	15.98718	18.30704	20.48318
11	2.60322	3.05348	3.81575	4.57481	5.57778	17.27501	19.67514	21.92005
12	3.07382	3.57057	4.40379	5.22603	6.30380	18.54935	21.02607	23.33666
13	3.56503	4.10692	5.00875	5.89186	7.04150	19.81193	22.36203	24.73560
14	4.07467	4.66043	5.62873	6.57063	7.78953	21.06414	23.68479	26.11895
15	4.60092	5.22935	6.26214	7.26094	8.54676	22.30713	24.99579	27.48839
16	5.14221	5.81221	6.90766	7.96165	9.31224	23.54183	26.29623	28.84535

Como α = 5% la región de aceptación de la hipótesis nula, es la región comprendida entre los valores: $[\chi^2_{97.5\%}, \chi^2_{2.5\%}] = [6.26, 27.49]$.

5. Conclusión

Como el valor observado o calculado está comprendido dentro de la región de aceptación de la H_0, se concluye con un nivel de 5% de significancia, que la máquina para empacar café tiene una varianza igual a 100 g^2, por lo tanto está bien regulada.

Ejemplo 96

La empresa de Autobuses Extraurbanos "Blanquita" hizo un esfuerzo para promover una imagen confiable, motivando a sus pilotos a mantener los horarios predeterminados de operación. Como política normal, la empresa espera que las horas de salida en diversas paradas tengan poca variabilidad. En términos de la varianza de los tiempos de salida, la norma de la empresa especifica una varianza de 4 o menos, con los tiempos en minutos.

En forma periódica se recopilan los datos de la hora de salida en diversas paradas, para determinar si se mantiene el lineamiento de variabilidad. Suponga que se obtiene una muestra aleatoria de 10 salidas de autobuses en determinada parada, que tiene una varianza $s^2 = 4.8$. ¿Es suficiente evidencia muestral para rechazar H_o y concluir que los autobuses no cumplen con el lineamiento de la varianza de tiempo de salida que establece la empresa? Utilice un nivel de 5% de significancia.

Solución

Datos: $n = 10$, $\sigma^2 = 4$ min^2, $s^2 = 4.8$ min^2, $\alpha = 0.05$

1. Establecer las hipótesis

H_0: $\sigma^2 \leq 4$ minutos^2 (la varianza de las horas de salida está dentro de los lineamientos de la empresa)

H_a: $\sigma^2 > 4$ minutos^2 (la varianza de las horas de salida no está dentro de los lineamientos de la empresa)

2. Cálculo de la estadística de prueba.

En este caso se utiliza la estadística χ^2 (ji cuadrada):

$$\chi^2_{o(n-1)} = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(10-1)4.8}{4} = 10.8$$

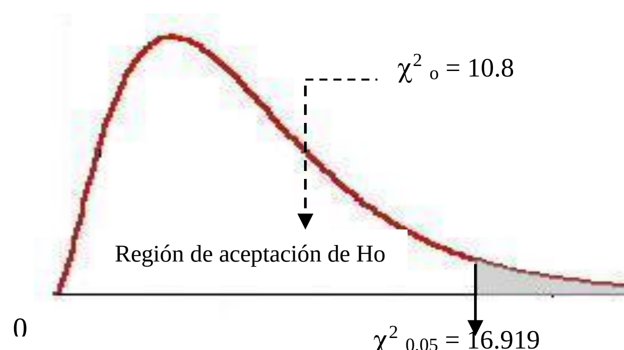
Quiere decir que el cociente arriba descrito tiene una distribución ji cuadrada con "n–1" grados de libertad.

3. Definir las regiones de rechazo y aceptación de Ho.

La tabla de ji cuadrada indica que, con $\alpha = 0.05$ (ya que es una prueba unilateral) y 9 grados de libertad, $\chi^2_{0.05} = 16.919$.

		Área correspondiente al extremo derecho de una distribución Jí-cuadrada.					
gl	**0.995**	**0.99**	**0.975**	**0.95**	**0.9**	**0.1**	**0.05**	**0.025**
1	0.00004	0.00016	0.00098	0.00393	0.01579	2.70554	3.84146	5.02389
2	0.01003	0.02010	0.05064	0.10259	0.21072	4.60517	5.99146	7.37776
3	0.07172	0.11483	0.21580	0.35185	0.58437	6.25139	7.81473	9.34840
4	0.20699	0.29711	0.48442	0.71072	1.06362	7.77944	9.48773	11.14329
5	0.41174	0.55430	0.83121	1.14548	1.61031	9.23636	11.07050	12.83250
6	0.67573	0.87209	1.23734	1.63538	2.20413	10.64464	12.59159	14.44938
7	0.98926	1.23904	1.68987	2.16735	2.83311	12.01704	14.06714	16.01276
8	1.34441	1.64650	2.17973	2.73264	3.48954	13.36157	15.50731	17.53455
9	1.73493	2.08790	2.70039	3.32511	4.16816	14.68366	16.91898	19.02277
10	2.15586	2.55821	3.24697	3.94030	4.86518	15.98718	18.30704	20.48318
11	2.60322	3.05348	3.81575	4.57481	5.57778	17.27501	19.67514	21.92005
12	3.07382	3.57057	4.40379	5.22603	6.30380	18.54935	21.02607	23.33666
13	3.56503	4.10692	5.00875	5.89186	7.04150	19.81193	22.36203	24.73560
14	4.07467	4.66043	5.62873	6.57063	7.78953	21.06414	23.68479	26.11895
15	4.60092	5.22935	6.26214	7.26094	8.54676	22.30713	24.99579	27.48839
16	5.14221	5.81221	6.90766	7.96165	9.31224	23.54183	26.29623	28.84535

La figura siguiente muestra la región de rechazo para esta prueba unilateral. Observe que sólo se rechaza H_o si la varianza s^2 de la muestra produce un valor grande de χ^2.



4. Regla de decisión.

Se rechaza la hipótesis nula H_0 sí el valor de χ^2 observado es mayor que χ^2_α (prueba unilateral a la derecha).

5. Conclusión

Como el valor observado o calculado está comprendido dentro de la región de aceptación de la H_0, la varianza de la muestra $s^2 = 4.8$ no constituye una evidencia que permita decir que la varianza de tiempos de salida no cumple con la norma de la empresa.

5.7 PRUEBA DE HIPÓTESIS ESTADÍSTICA ACERCA DE DOS MEDIAS POBLACIONALES INDEPENDIENTES, CON VARIANZAS DESCONOCIDAS E IGUALES.

Ejemplo 97

Los datos de la siguiente tabla se refieren a las alturas (en metros) de árboles en muestras aleatorias e independientes de dos especies forestales diferentes (1 y 2). Verifique si las alturas medias de los árboles de las dos especies no difieren entre sí, considerándose un nivel de significancia del 5%.

Especie 1					**Especie 2**		
23.4	24.4	24.6	24.9	22.5	22.9	23.7	24.0
25.0	26.2	26.3	26.8	24.4	24.5	25.3	26.0
26.8	26.9	27.0	27.6	26.2	26.4	26.7	26.9
27.7				27.4	28.5		

Solución:

Vamos a suponer que las dos poblaciones tengan la misma variancia $\sigma_1^2 = \sigma_2^2$, sin embargo son desconocidas.

Datos:

	Especie 1	Especie 2
Media	25.97	25.39
Desviación estándar	1.36	1.77
Tamaño de la muestra	13	14

1. Establecer las hipótesis

H_0: $\mu_1 = \mu_2$ (las alturas promedio de los árboles de las dos especies son iguales)
H_a: $\mu_1 \neq \mu_2$ (las alturas promedio de los árboles de las dos especies son diferentes)

2. Cálculo de la estadística de prueba.

En este caso se utiliza la estadística t de Student. Para obtener el valor observado o calculado de t se utiliza la ecuación siguiente:

$$t_o = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

- Error estándar estimado de la diferencia entre dos medias muestrales

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Estimación conjunta de σ^2 (varianza combinada)

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = s_p^2 = \frac{(13-1)1.36^2+(14-1)1.77^2}{13+14-2} = \frac{62.92}{25} = 2.52$$

$s_p = \sqrt{s_p^2} = \sqrt{2.52} = 1.59$, por lo tanto:

$$\hat{\sigma}_{\bar{x}_1-\bar{x}_2} = 1.59\sqrt{\frac{1}{13}+\frac{1}{14}} = (1.59)(0.39) = 0.62 \text{, entonces:}$$

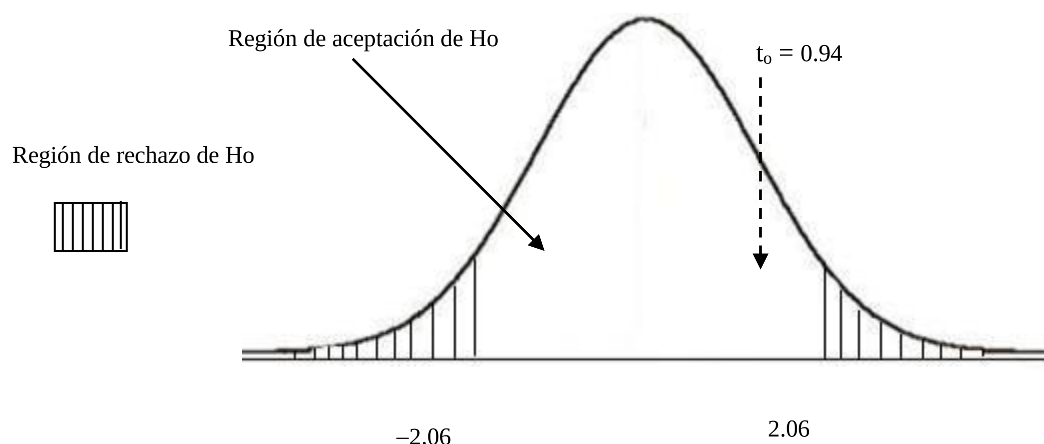
$$t_o = \frac{\bar{x}_1-\bar{x}_2}{\hat{\sigma}_{\bar{x}_1-\bar{x}_2}} = \frac{25.97-25.39}{0.62} = 0.94$$

3. Definir las regiones de rechazo y aceptación de Ho.

El valor crítico de t, que separa a las regiones de rechazo y de aceptación de la hipótesis nula, está en función de: n_1+n_2-2 grados de libertad y un determinado nivel de significancia (α). Para este caso, tenemos que buscar en la tabla t de Student, con 25 grados de libertad y α = 0.05 (con dos colas)

Tabla t de Student

	0.10	0.05	0.025	0.01	0.005	Una cola
gl	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**Dos colas**
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
21	1.323	1.721	2.080	2.518	2.831	
22	1.321	1.717	2.074	2.508	2.819	
23	1.319	1.714	2.069	2.500	2.807	
24	1.318	1.711	2.064	2.492	2.797	
25	1.316	1.708	2.060	2.485	2.787	
26	1.315	1.706	2.056	2.479	2.779	



4. Regla de decisión

Así, fijando el nivel de significancia "α", la hipótesis nula será rechazada sí:
El valor absoluto de t_o es mayor que el valor de $t_{\alpha/2}$ en la prueba bilateral; o bien,
$t_o > t_\alpha$, en la prueba unilateral a la derecha y
$t_o < t_\alpha$ en la prueba unilateral a la izquierda.

5. Conclusión

Debido a que el valor de t observado (0.94) es menor que el valor crítico de t (2.060) se concluye que, con un nivel de significancia del 5%, las alturas medias de los árboles de las dos especies no difieren entre sí.

Resolución del ejemplo 97 con Infostat

1. Ingreso de datos

Caso	especie	altura
1	especie 1	23.4
2	especie 1	24.4
3	especie 1	24.6
4	especie 1	24.9
5	especie 1	25.0
6	especie 1	26.2
7	especie 1	26.3
8	especie 1	26.8
9	especie 1	26.8
10	especie 1	26.9
11	especie 1	27.0
12	especie 1	27.6
13	especie 1	27.7
14	especie 2	22.5
15	especie 2	22.9
16	especie 2	23.7
17	especie 2	24.0
18	especie 2	24.4
19	especie 2	24.5
20	especie 2	25.3
21	especie 2	26.0
22	especie 2	26.2
23	especie 2	26.4
24	especie 2	26.7

2. Solicitar análisis

Prueba T para muestras Independientes

prueba
Bilateral
Unilateral derecha
Unilateral Izquierda

Comparaciones
Todos contra todos
Selección vs restantes

N obs. Medias
Varianzas p
gl T
Inter conf.

Grupos activos
especie 1
especie 2

Grupos omitidos

Si las varianzas no son homogéneas a nivel del 5 %
hacer automáticamente una prueba T para varianzas heterogéneas

Aceptar Cancelar Ayuda

3. Verificar resultados y comparar con los cálculos manuales

Prueba T para muestras Independientes

Clasific	Variable	Grupo(1)	Grupo(2)	n(1)	n(2)	media(1)	media(2)	p(Var.Hom.)	T	gl	p	prueba
especie	altura	{especie 1}	{especie 2}	13	14	25.97	25.39	0.3647	0.95	25	0.3491	Bilateral

Ejemplo 98

Un nuevo programa de cómputo ha sido desarrollado por la empresa MILLENIUM para ayudar a los analistas de sistemas a reducir el tiempo requerido para diseñar, desarrollar e implementar un sistema de información. Para evaluar las ventajas del programa, se selecciona una muestra aleatoria de 24 analistas de sistemas. A cada analista se le proporcionan especificaciones para un sistema hipotético de información, y a 12 de ellos se les pide producir el sistema usando la tecnología (los programas) actual. A los otros 12 se les capacita primero en el uso del nuevo paquete y, a continuación, se les pide usarlo para producir el sistema de información.

El investigador a cargo del proyecto de evaluación del nuevo programa espera demostrar que ese paquete permite un menor tiempo promedio de terminación del proyecto. El resumen de los resultados de la prueba se presenta a continuación:

	Tecnología actual Grupo 1	Nuevo programa Grupo 2
Media	$\bar{x}_1 = 325\ horas$	$\bar{x}_2 = 288\ horas$
Desviación estándar	$s_1 = 40\ horas$	$s_2 = 44\ horas$
Tamaño de la muestra	$n_1 = 12$	$n_2 = 12$

Solución:
Vamos a suponer que las dos poblaciones tengan la misma variancia $\sigma_1^2 = \sigma_2^2$, sin embargo son desconocidas.

1. Establecer las hipótesis

H_0: $\mu_1 - \mu_2 \leq 0$ (la diferencia entre las dos medias poblacionales es menor o igual que cero)
H_a: $\mu_1 - \mu_2 > 0$ (la diferencia entre las dos medias poblacionales es mayor que cero)

2. Cálculo de la estadística de prueba.

En este caso se utiliza la estadística t de Student. Para obtener el valor observado o calculado de t se utiliza la ecuación siguiente (tanto para n diferentes como iguales):

$$t_o = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

$$t_o = \frac{325 - 288}{\sqrt{(12-1)40^2 + (12-1)44^2}} \sqrt{\frac{12*12(12+12-2)}{12+12}} = 2.155$$

O bien la ecuación cuando los tamaños de las muestras son iguales:

$$t_o = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{40^2}{12} + \frac{44^2}{12}} = \sqrt{133.33 + 161.33} = 17.17$$

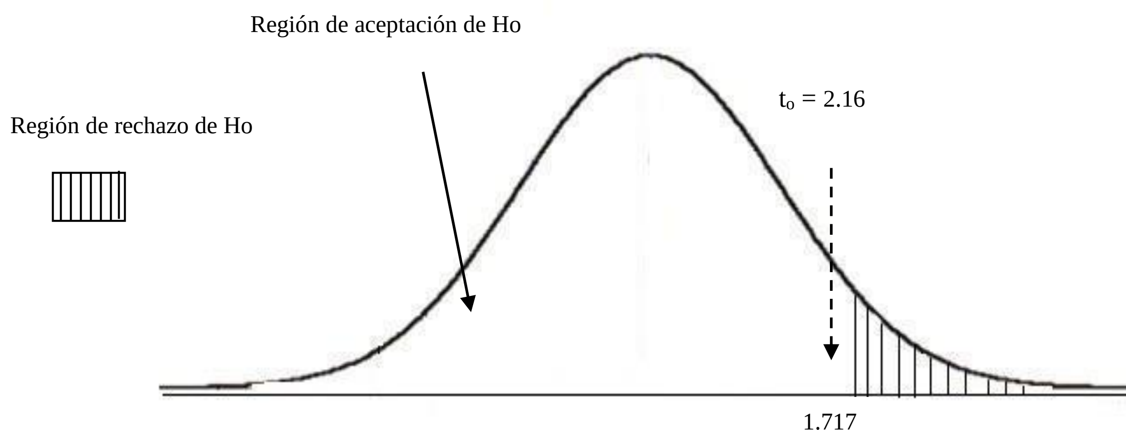
$$z_o = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} = \frac{14.3 - 15.7}{0.706} = \frac{-1.4}{0.706} = -1.982$$

3. Definir las regiones de rechazo y aceptación de Ho.

El valor crítico de t, que separa a las regiones de rechazo y de aceptación de la hipótesis nula, está en función de: n_1+n_2-2 grados de libertad y un determinado nivel de significancia (α). Para este caso, tenemos que buscar en la tabla t de Student, con 22 grados de libertad y α = 0.05 (para una cola, por tratarse de una prueba unilateral)

Tabla t de Student

	0.10	0.05	0.025	0.01	0.005	**Una cola**
gl	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**Dos colas**
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
21	1.323	1.721	2.080	2.518	2.831	
22	1.321	1.717	2.074	2.508	2.819	
23	1.319	1.714	2.069	2.500	2.807	
24	1.318	1.711	2.064	2.492	2.797	
25	1.316	1.708	2.060	2.485	2.787	
26	1.315	1.706	2.056	2.479	2.779	



4. Regla de decisión

Así, fijando el nivel de significancia "α", la hipótesis nula será rechazada sí:
El valor de $t_o > t_\alpha$, en la prueba unilateral a la derecha.

5. Conclusión

Debido a que el valor de t observado (2.16) es mayor que el valor crítico de t (1.717) se concluye que, con un nivel de significancia del 5%, $\mu_1 - \mu_2 > 0$ y que el nuevo programa de cómputo si permite menores tiempos promedio de terminación.

5.8 PRUEBA DE HIPÓTESIS ESTADÍSTICA ACERCA DE DOS MEDIAS POBLACIONALES INDEPENDIENTES, PROVENIENTES DE MUESTRAS GRANDES.

Ejemplo 99

Se compararon dos marcas de cigarrillos, Alas y Strike, respecto a su contenido de nicotina en miligramos, dieron los siguientes resultados.

Alas	**Strike**
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = 14.3$	$\bar{x}_2 = 15.7$
$s_1 = 2.9$	$s_2 = 3.8$

Con un nivel de significancia de 0.01. Existe suficiente evidencia estadística para decir que hay diferencia entre las medias de contenido de nicotina para las dos marcas de cigarrillos.

1. Establecer las hipótesis

H_0: $\mu_1 = \mu_2$ (los contenidos promedios de nicotina son iguales en las dos marcas de cigarros)
H_a: $\mu_1 \neq \mu_2$ (los contenidos promedios de nicotina son diferentes en las dos marcas de cigarros)

2. Cálculo de la estadística de prueba.

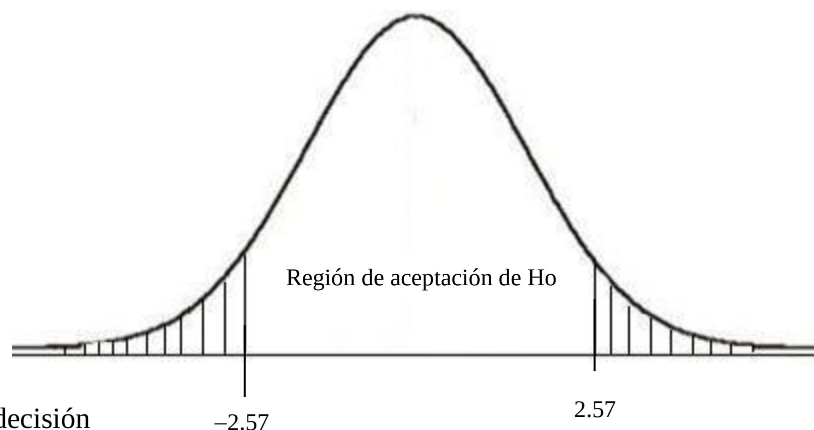
En este caso se utiliza la estadística z de la distribución normal. Para obtener el valor observado o calculado de z se utiliza la ecuación siguiente:

$$z_o = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

$$z_o = \frac{14.3 - 15.7}{\sqrt{(40-1)2.9^2 + (50-1)3.8^2}} \sqrt{\frac{40*50(40+50-2)}{40+50}} = -1.924$$

3. Definir las regiones de rechazo y aceptación de Ho.

Los valores críticos de z, para un nivel de significancia del 1%, utilizando una prueba bilateral son –2.57 y 2.57.



4. Regla de decisión

Se acepta la hipótesis nula H_0 sí el valor de z_o está comprendido dentro del rango −2.57 y 2.57.

5. Conclusión

Debido a que el valor absoluto de z observado (1.924) no es mayor que el valor crítico de z (2.57) se concluye que, con un nivel de significancia del 1%, no existe suficiente evidencia estadística para decir que hay diferencia entre las medias de contenido de nicotina para las dos marcas de cigarrillos.

5.9 COMPARACIÓN DE MEDIAS INDEPENDIENTES CON VARIANZAS DESCONOCIDAS (prueba de Welch)

Como las varianzas son desconocidas es necesario estimarlas a través de las varianzas muestrales S_X^2 y S_Y^2. En este caso, al substituir las varianzas poblacionales por las muestrales en la expresión:

$$\frac{\overline{X}-\overline{Y}-\Delta}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}$$

se obtiene la estadística:

$$t = \frac{\overline{X}-\overline{Y}-\Delta}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}}$$

No se tendrá una distribución normal, pero si una distribución "t" con los grados de libertad corregidos por la siguiente expresión (Welch–Satterthwaite):

$$\nu = \frac{\left(\frac{S_X^2}{n}+\frac{S_Y^2}{m}\right)^2}{\frac{\left(\frac{S_X^2}{n}\right)^2}{n-1}+\frac{\left(\frac{S_Y^2}{m}\right)^2}{m-1}}$$

Siempre y cuando (n, m) sean mayores o iguales a 30, o las muestras hayan sido extraídas de poblaciones que tengan distribuciones normales.

Así, fijando el nivel de significancia "α", la hipótesis nula será rechazada sí:
$|t_c| > t_{\alpha/2}$ en la prueba bilateral;
$t_c > t_\alpha$, en la prueba unilateral a la derecha y
$t < t_\alpha$ en la prueba unilateral a la izquierda,

Ejemplo 100

Las resistencias de dos tipos de concreto fueron medidas, los resultados se muestran en la siguiente tabla. Fijando un nivel de significancia de 5%, ¿existe evidencia de que el concreto del tipo A sea más resistente que el concreto del tipo B?

Tipo *A*	54	55	58	51	57
Tipo *B*	50	54	56	52	53

Solución:
Las hipótesis son:
H_0: $\mu_A - \mu_B = 0$ ($\mu_A = \mu_B$) contra
H_a: $\mu_A - \mu_B > 0$ ($\mu_A > \mu_B$)
Los datos obtenidos de la tabla son:

$\bar{x} = 55.0$ y $\bar{y} = 53.0$
$S_X^2 = 7.50$ y $S_Y^2 = 5.0$

El valor de la estadística observada será:

$$t_o = \frac{55 - 53}{\sqrt{\frac{7.5}{5} + \frac{5.0}{5}}} = 1.265$$

Con $\alpha = 5\%$, y los grados de libertad (v):

$$\nu = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{\left(\frac{S_X^2}{n}\right)^2}{n-1} + \frac{\left(\frac{S_Y^2}{m}\right)^2}{m-1}} = \frac{\left(\frac{7.5}{5} + \frac{5}{5}\right)^2}{\frac{\left(\frac{7.5}{5}\right)^2}{4} + \frac{\left(\frac{5}{5}\right)^2}{4}} = \frac{6.25}{0.8125} = 7.69 \cong 8,$$

entonces el valor de "t" crítico con 8 grados libertad será: 1.86. En este caso, con estas muestras no es posible afirmar que el concreto del tipo A sea más resistente el concreto del tipo B.

5.10 PRUEBA DE HIPÓTESIS ESTADÍSTICA ACERCA DE DOS MEDIAS POBLACIONALES DEPENDIENTES (O PAREADAS)

Ejemplo 101

Fue impartió un curso sobre “Atención al cliente” a 10 empleados de un ingenio azucarero. Para evaluar el curso, se realizó una prueba antes y otra después de habérseles impartido la capacitación a los empleados. Pruebe con un nivel de significancia de 0.10 si existe evidencia para decir que la media de la diferencia en los puntajes antes y después de la capacitación es diferente. Los puntajes aparecen en la tabla siguiente:

Solución

1. Establecer la hipótesis

Ho: $\mu_D = \mu_A - \mu_D = 0$ (la media de las diferencias de los punteos es igual a cero)

Ha:

Empleado	Puntaje antes de la capacitación del empleado	Puntaje después de la capacitación del empleado	Diferencia (d_i)	$(d_i - \bar{d})^2$
1	9.00	9.20	-0.20	0.04
2	7.30	8.20	-0.90	0.25
3	6.70	8.50	-1.80	1.96
4	5.30	4.90	0.40	0.64
5	8.70	8.90	-0.20	0.04
6	6.30	5.80	0.50	0.81
7	7.90	8.20	-0.30	0.01
8	7.30	7.80	-0.50	0.01
9	8.00	9.50	-1.50	1.21
10	8.50	8.00	0.50	0.81
		Sumatoria	-4.00	5.78
		Promedio ($\bar{d}$)	-0.40	

$\mu_D = \mu_A - \mu_D \neq 0$ (la media de las diferencias de los punteos es diferente de cero)

2. Cálculo de la estadística de prueba.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{-4.0}{10} = -0.40 \qquad S_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{5.78}{9}} = 0.801$$

Con esta información, se procede a calcular el estadístico t:

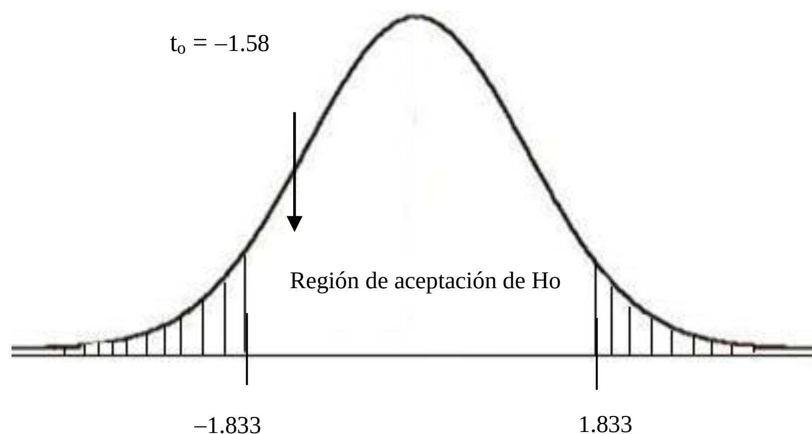
$$t_o = \frac{\bar{d} - \mu_D}{S_{\bar{d}} / \sqrt{n}} \qquad t_o = \frac{-0.4 - 0}{0.801 / \sqrt{10}} = -1.58$$

3. Definir nivel de significancia y zona de rechazo.

Con un nivel de significancia = 0.10 para una prueba bilateral y 9 grados de libertad, tenemos que el valor crítico es:

Tabla t de Student						
	0.10	0.05	0.025	0.01	0.005	**Una cola**
gl	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**Dos colas**
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
4	1.533	2.132	2.776	3.747	4.604	
5	1.476	2.015	2.571	3.365	4.032	
6	1.440	1.943	2.447	3.143	3.707	
7	1.415	1.895	2.365	2.998	3.499	
8	1.397	1.860	2.306	2.896	3.355	
9	1.383	1.833	2.262	2.821	3.250	
10	1.372	1.812	2.228	2.764	3.169	
11	1.363	1.796	2.201	2.718	3.106	
12	1.356	1.782	2.179	2.681	3.055	
13	1.350	1.771	2.160	2.650	3.012	
14	1.345	1.761	2.145	2.624	2.977	
15	1.341	1.753	2.131	2.602	2.947	

Gráficamente se tiene:



4. Conclusión

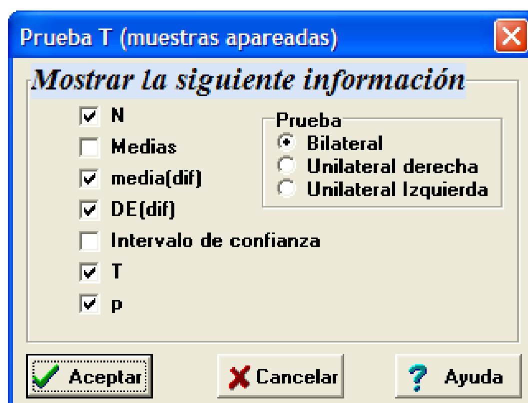
Como el valor de to está comprendido dentro de la región de aceptación de la hipótesis nula, se concluye que estadísticamente la media de la diferencia en los puntajes antes y después de la capacitación es igual, por lo que no hay evidencia para indicar que la capacitación haya tenido efecto.

Resolución del Ejemplo 101 con Infostat

1. Ingreso de datos

Caso	antes	después
1	9.00	9.20
2	7.30	8.20
3	6.70	8.50
4	5.30	4.90
5	8.70	8.90
6	6.30	5.80
7	7.90	8.20
8	7.30	7.80
9	8.00	9.50
10	8.50	8.00

2. Solicitar análisis



3. Verificar resultados y comparar con los cálculos manuales

Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	DE(dif)	T	Bilateral
antes	después	10	-0.40	0.80	-1.58	0.1489

Ejemplo 102

Con frecuencia, los procesadores de palabras se justifican con base en una mayor eficiencia del personal secretarial. A continuación vemos rapidez de mecanografiado, en palabras por minuto de 7 secretarias que antes usaban máquinas de escribir electrónicas, y que hoy usan el procesador de palabras StarWord®. Pruebe, con un nivel de significancia de 0.05, si aumentó la media de la rapidez de mecanografiado con el procesador de palabras.

Secretaria	Máquina de escribir eléctrica (ME)	Procesador de palabras (PP)	Diferencia (d_i)	$(d_i - \bar{d})^2$
1	72	75	-3	0.32
2	68	66	2	19.62
3	55	60	-5	6.60
4	58	64	-6	12.74
5	52	55	-3	0.32
6	55	57	-2	0.18
7	64	64	0	5.90
		Sumatoria	-17	45.71
		Promedio ($\bar{d}$)	-2.43	

Solución

1. Establecer la hipótesis

Ho: $\mu_{ME} \geq \mu_{PP}$ (la media de la rapidez de mecanografiado con máquina eléctrica es superior a la del procesador de texto)

Ha: $\mu_{ME} < \mu_{PP}$ (la media de la rapidez de mecanografiado con máquina eléctrica es inferior a la del procesador de texto)

2. Cálculo de la estadística de prueba.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{-17}{7} = -2.43 \qquad S_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{45.71}{6}} = 2.76$$

Con esta información, se procede a calcular el estadístico t:

$$t_o = \frac{\bar{d} - \mu_D}{S_{\bar{d}} / \sqrt{n}} \qquad t_o = \frac{-2.43 - 0}{2.76 / \sqrt{7}} = -2.33$$

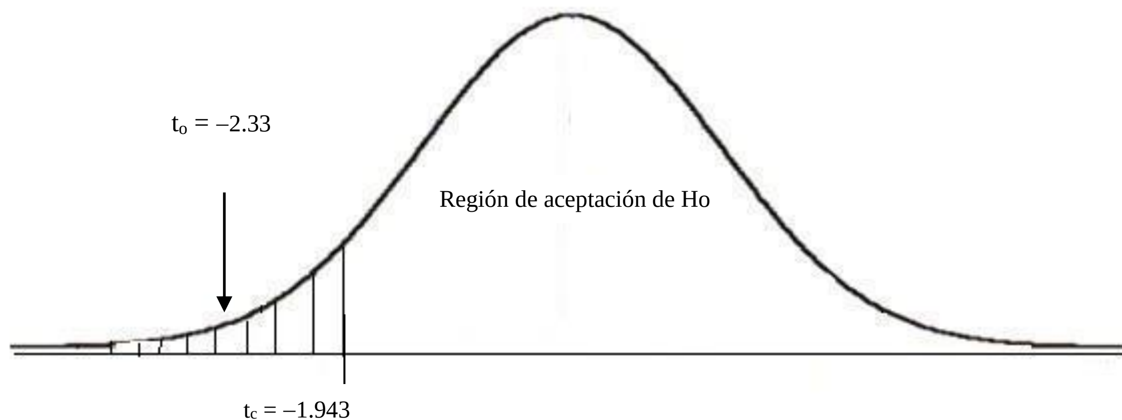
3. Definir nivel de significancia y zona de rechazo.

Con un nivel de significancia = 0.05 para una prueba unilateral y 6 grados de libertad, tenemos que el valor crítico es:

Tabla t de Student

	0.10	0.05	0.025	0.01	0.005	**Una cola**
gl	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**Dos colas**
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
4	1.533	2.132	2.776	3.747	4.604	
5	1.476	2.015	2.571	3.365	4.032	
6	1.440	1.943	2.447	3.143	3.707	
7	1.415	1.895	2.365	2.998	3.499	
8	1.397	1.860	2.306	2.896	3.355	
9	1.383	1.833	2.262	2.821	3.250	
10	1.372	1.812	2.228	2.764	3.169	
11	1.363	1.796	2.201	2.718	3.106	
12	1.356	1.782	2.179	2.681	3.055	
13	1.350	1.771	2.160	2.650	3.012	
14	1.345	1.761	2.145	2.624	2.977	
15	1.341	1.753	2.131	2.602	2.947	

Gráficamente se tiene:



4. Conclusión

Como el valor de to está comprendido dentro de la región de rechazo de la hipótesis nula, se concluye que estadísticamente aumentó la media de la rapidez de mecanografiado con el procesador de palabras.

5.11 PRUEBA DE HIPÓTESIS ESTADÍSTICA PARA LA COMPARACIÓN DE LAS VARIANZAS DE DOS POBLACIONES NORMALES.

Ejemplo 103

Un fabricante de productos farmacéuticos compra cierto material de dos proveedores. El contenido medio de impurezas en la materia prima es aproximadamente el mismo para ambos proveedores, pero el fabricante está preocupado por la variación de las impurezas de un embarque a otro.
Si el contenido de impurezas tendiera a variar excesivamente con respecto a una fuente de suministro, podría afectar la calidad del producto farmacéutico. Para comparar la variación en el porcentaje de impurezas para los dos proveedores, el fabricante selecciona al azar diez embarques de cada uno de los proveedores y mide el porcentaje de impurezas en la materia prima de cada embarque. Las medias muestrales y las varianzas se muestran en la siguiente tabla.

	Proveedor A	Proveedor B
Media muestral	1.89	1.85
Varianza muestral	0.273	0.094

¿Proporcionan los datos evidencia suficiente para indicar una diferencia en la variación de los contenidos de impurezas para los embarques de los dos proveedores? Realice la prueba con un nivel de significancia del 5%

Solución:

1. Hipótesis

H_o: $\sigma_A^2 = \sigma_B^2$ (la varianza del porcentaje de impurezas es igual en la materia prima de los dos proveedores)

H_a: $\sigma_A^2 \neq \sigma_B^2$ (la varianza del porcentaje de impurezas es distinta en la materia prima de los dos proveedores)

3. Estadístico calculado u observado

$$F_{obs} = \frac{S_{Mayor}{}^{2}}{S_{Menor}{}^{2}} = \frac{0.273}{0.094} = 2.9042$$

Nota: Se coloca el mayor valor de varianza en el denominador, con la finalidad de poder trabajar con la cola superior de la distribución de F.

4. Valor crítico de la estadística

En este caso, se utilizará la distribución F de Fisher-Snedecor, con n_1-1 grados de libertad en el numerador, y n_2-1 grados de libertad en el denominador. Como tenemos 10 embarques para cada proveedor, se tiene que $n_1 = n_2 = 9$ grados de libertad.

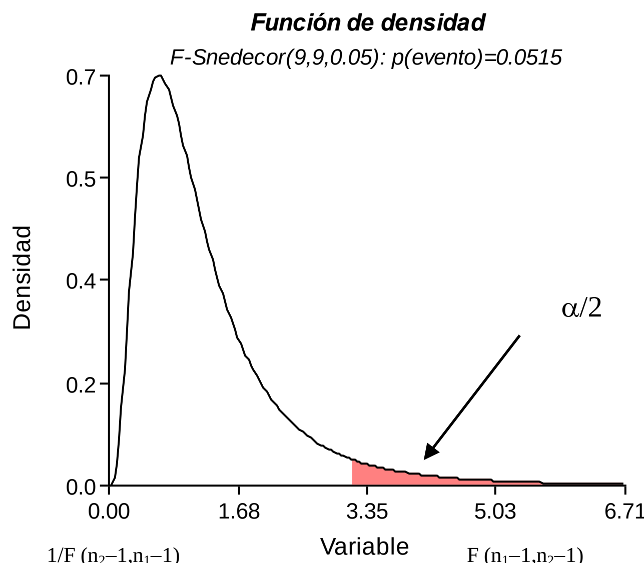
TABLA F, NIVEL DE SIGNIFICANCIA = 5%

Grados de libertad	Grados de libertad del numerador																			
denominador	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	242,98	243,91	244,69	245,36	245,95	246,46	246,92	247,32	247,69	248,01
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,42	19,42	19,43	19,43	19,44	19,44	19,44	19,45
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,73	8,71	8,70	8,69	8,68	8,67	8,67	8,66
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,89	5,87	5,86	5,84	5,83	5,82	5,81	5,80
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,60	4,59	4,58	4,57	4,56
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,92	3,91	3,90	3,88	3,87
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51	3,49	3,48	3,47	3,46	3,44
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22	3,20	3,19	3,17	3,16	3,15
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01	2,99	2,97	2,96	2,95	2,94
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85	2,83	2,81	2,80	2,79	2,77
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72	2,70	2,69	2,67	2,66	2,65

El valor de F crítico es: $F_{(9,9,0.05)} = 3.18$

5. Región de rechazo y de aceptación

Luego de planteadas las hipótesis, estimado el valor de F y definido el nivel de significancia, se establece la región crítica (RC) de la prueba de la siguiente manera:



6. Conclusión:

Como el valor de la estadística F =2.9042 es menor al valor crítico de la estadística F =3.18, se acepta la Ho, por lo tanto se puede decir que no existe evidencia suficiente para indicar una diferencia en la variación de los contenidos de impurezas para los embarques de los dos proveedores.

5.12 PRUEBA DE HIPÓTESIS ESTADÍSTICA ACERCA DE LA DIFERENCIA ENTRE LAS PROPORCIONES DE DOS POBLACIONES NORMALES.

Ejemplo 104

La empresa VOXSTAT realiza entrevistas de puerta en puerta sobre una diversidad de asuntos. Algunas personas cooperan con el entrevistador y llenan el cuestionario, y otras no. Se dispone de los siguientes datos:

Encuestados	Tamaño de la muestra	No. de personas que cooperan
1. Hombres	200	110
2. Mujeres	300	210

Pruebe la hipótesis que la tasa de respuestas es igual para hombres y mujeres, con α = 0.05.

Solución:

Datos: $\bar{p}_1 = \dfrac{110}{200} = 0.55$ $\bar{p}_2 = \dfrac{210}{300} = 0.70$

1. Establecer las hipótesis

H_0: $p_1 = p_2$ (no hay diferencia significativa entre las proporciones de las dos poblaciones)
H_a: $p_1 \neq p_2$ (la proporción de hombres que responden el cuestionario es diferentes a la proporción de mujeres)

2. Cálculo de la estadística de prueba.

Primero se deben combinar las dos proporciones para obtener un estimado. Este estimador combinado, representado con $\bar{p}_c$, es el siguiente:

$$\bar{p}_c = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}, \quad \bar{p}_c = \frac{(200)(0.55) + (300)(0.70)}{200 + 300} = 0.64$$

En este caso se utiliza la estadística z de la distribución normal. Para obtener el valor observado o calculado de z se utiliza la ecuación siguiente:

$$z_o = \frac{\bar{p}_1 - \bar{p}_2}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}}$$

- Error estándar estimado de la diferencia entre dos proporciones muestrales:

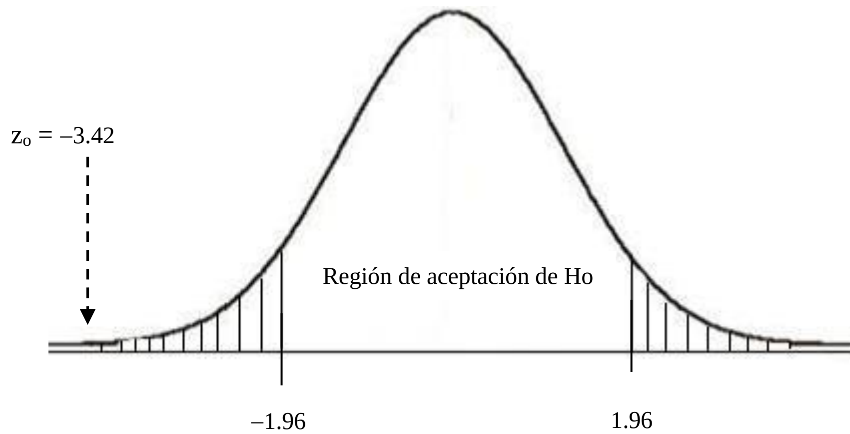
$$\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\bar{p}_c (1 - \bar{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.64(0.36)\left(\frac{1}{200} + \frac{1}{300}\right)} = \sqrt{0.2304\,(0.0083)} = 0.04382$$

- Estadístico de prueba:

$$z_o = \frac{\bar{p}_1 - \bar{p}_2}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}} = = \frac{0.55 - 0.70}{0.04382} = \frac{-0.15}{0.04382} = -3.42$$

3. Definir las regiones de rechazo y aceptación de Ho.

Los valores críticos de z, para un nivel de significancia del 5%, utilizando una prueba bilateral son –1.96 y 1.96. Tal como se ilustra en la gráfica siguiente:



4. Regla de decisión

Se acepta la hipótesis nula H_0 sí el valor de z_o está comprendido dentro del rango –1.96 y 1.96.

5. Conclusión

Debido a que el valor de z observado (–3.42) está en la región de rechazo de la hipótesis nula, se concluye que, con un nivel de significancia del 5%, que existe diferencia en la proporción de respuestas para hombres y mujeres.

5.13 PRUEBA χ^2 DE INDEPENDENCIA.

Si se dispone de la función de probabilidad conjunta de dos variables aleatorias, se puede verificar que, para todos los posibles valores de las variables, el producto de las probabilidades marginales es igual a la probabilidad conjunta. En la situación más común en la que no se tiene información sobre la ocurrencia conjunta de las variables aleatorias, el procedimiento usual es colectar una muestra anotando la frecuencia conjunta de la ocurrencia de los valores de las variables. Se puede, entonces, utilizar una prueba de hipótesis conocida como: prueba de independencia.

Procedimiento.

1. Ho: No existe relación entre las dos variables (las variables son independientes).
   Ha: Si existe relación entre las dos variables (las variables no son independientes).

2. Cálculo de las frecuencias esperadas. Se debe construir una tabla de contingencia k × r, siendo k = número de columnas y r = número de filas. Ejemplo de una tabla 2 × 2.

A	B	A + B= $\eta_{1.}$
C	D	C + D= $\eta_{2.}$
A +C= $\eta_{.1}$	B + D = $\eta_{.2}$	

$$e_{ij} = \frac{Total\ de\ la\ línea\ i \times Total\ de\ la\ línea\ j}{Total\ general} = \frac{\eta_{i.} \times \eta_{.j}}{\eta}$$

Note que los valores esperados son calculados bajo la hipótesis Ho de independencia y, por esa razón, se utilizan los totales de línea y columna que representan las frecuencias marginales de las variables.

3. Cálculo del estadístico de prueba.

$\chi_o^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$, o bien, cuando se tiene un grado de libertad, se utiliza la corrección de Yates: $\chi_o^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(|o_{ij}-e_{ij}|-0.5)^2}{e_{ij}}$.

4. Cálculo del número de grados de libertad, .gl = (k – 1) (r – 1).

5. Regla de decisión.

Se rechaza Ho si $\chi_o^2 \geq \chi^2_{\text{crítica } (\alpha/2)}$ o bien si $\chi_o^2 \leq \chi^2_{\text{crítica } (1-\alpha/2)}$

Ejemplo 105

En una comunidad se realizó un estudio socioeconómico. Con los datos de tenencia de la tierra y emigración temporal se construyó la siguiente tabla de contingencia (o de doble entrada). Se desea saber si existe relación entre emigración temporal y tenencia de la tierra.

Posee tierra propia	Emigra		
	Si	No	
Si	45	23	68
No	51	28	79
	96	51	147

Solución:

1. Ho: La emigración temporal es independiente de la tenencia de la tierra.
   Ha: La emigración temporal no es independiente de la tenencia de la tierra.

2. Cálculo de las frecuencias esperadas.

   Debido a que gl = (2–1) (2–1) = 1, se utilizará la corrección de Yates, para el cálculo de las frecuencias esperadas.

$$\chi_o^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(|o_{ij}-e_{ij}|-0.5)^2}{e_{ij}} =$$

$$\frac{(|45-44.1|-0.5)^2}{44.1}+\frac{(|23-23.59|-0.5)^2}{23.59}+\frac{(|51-51.59|-0.5)^2}{51.59}\frac{(|28-27.41|-0.5)^2}{27.41}=0.0009$$

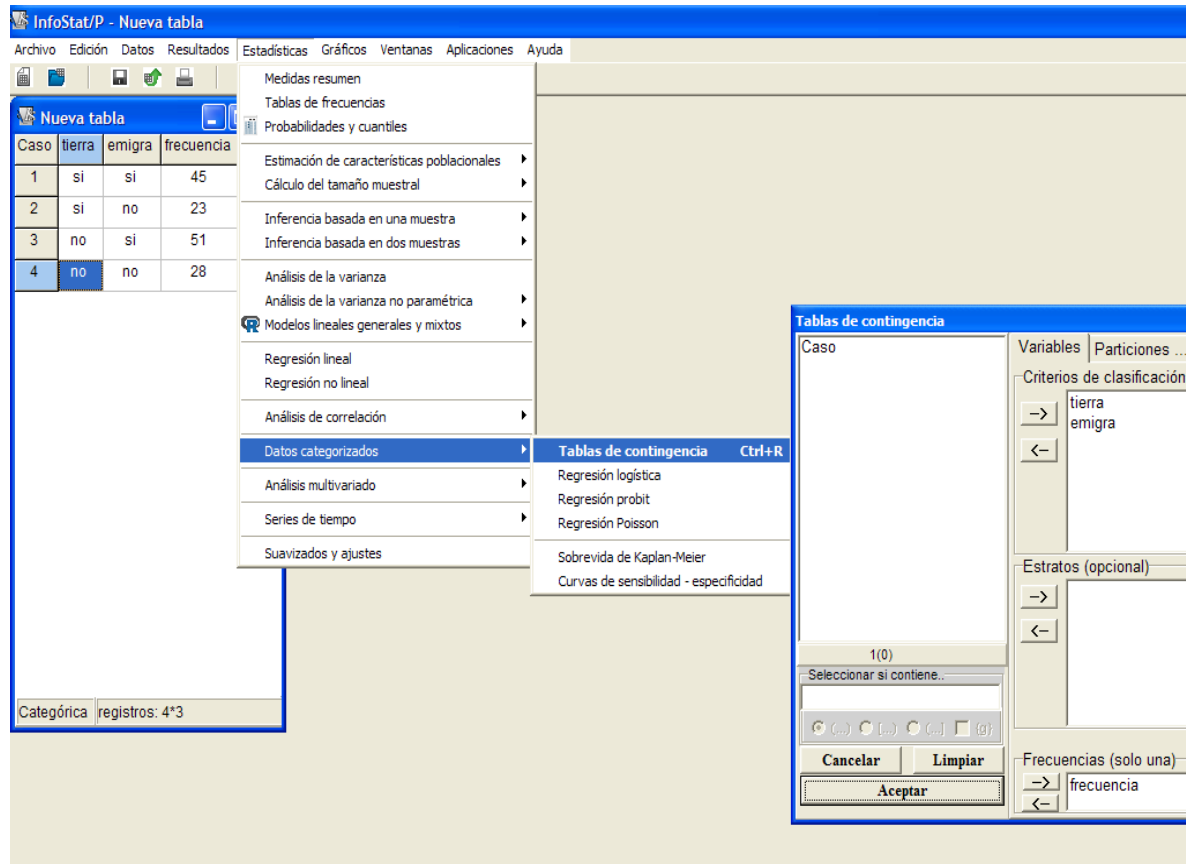
3. Región crítica. Debido a que el valor de χ_o^2 es pequeño, se utilizará la cola inferior de la distribución Jí-cuadrada. Entonces $\chi^2_{\text{crítica } (0.95,\ 1)} = 0.0039$

4. Regla de decisión: Como se está utilizando la cola inferior y $\chi_o^2 > \chi^2_{\text{crítica}\ (0.95,\ 1)}$, se acepta Ho.

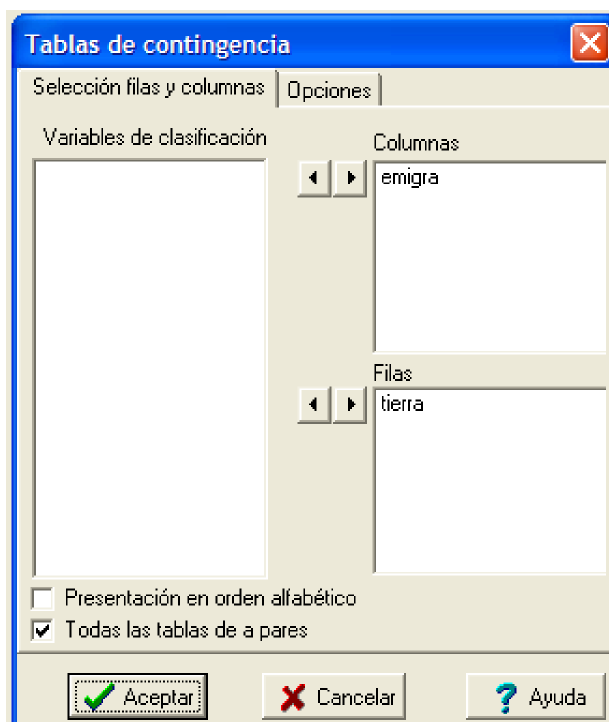
5. Conclusión: La tenencia de la tierra y la emigración son independientes.

Resolución de ejercicio 105 con Infostat

1. Ingreso de datos



2. Solicitar análisis



3. Verificar resultados y comparar con los cálculos manuales

Tablas de contingencia

Frecuencias: frecuencia

Frecuencias absolutas

En columnas:emigra

tierra	si	no	Total
si	45	23	68
no	51	28	79
Total	96	51	147

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	0.04	1	0.8370
Chi Cuadrado MV-G2	0.04	1	0.8370
Irwin-Fisher bilateral	0.02		0.8636
Coef.Conting.Cramer	0.01		
Coef.Conting.Pearson	0.02		
Coeficiente Phi	0.02		

5.14 PRUEBA χ^2 DE BONDAD DE AJUSTE

Ejemplo 106

Un vendedor de la compañía Forest Paper, tiene que visitar **cinco clientes por día**. Se sugiere que la variable medida por el vendedor puede ser descrita mediante una distribución binomial, con una probabilidad de éxito en cada visita de **p = 0.4**. Dada la siguiente distribución de frecuencias del número de ventas por día realizadas por Armstrong, ¿se puede concluir que los datos, de hecho, siguen la distribución sugerida? Utilice un nivel de 5% de significancia.

No. ventas diarias	0	1	2	3	4	5
Frecuencia del número de ventas	10	41	60	20	6	3

Solución:

1. Ho : X ~ B (5,0.4). La frecuencia del número de ventas por día sigue una distribución binomial.
   Ha: La frecuencia del número de ventas por día no sigue una distribución binomial.

2. Cálculo de las frecuencias esperadas.

Categorías	No. ventas diarias (x)	Frecuencia observada (o_i)	Probabilidad binomial P (X=x)	Total de clientes visitados	Frecuencia esperada (e_i)
1	0	10	0.078	140	10.92
2	1	41	0.259	140	36.26
3	2	60	0.346	140	48.44
4	3	20	0.230	140	32.20
5	4	6	0.077	140	10.78
6	5	3	0.010	140	1.40
		140			

$e_i = n \times p_i$, por ejemplo: $e_1 = 140 \times 0.078 = 10.92$

** Nota: la categoría 6 se unirá a la 5.

3. Cálculo de la estadística Jí-cuadrada.

$$\chi_o^2 = \frac{(10-10.92)^2}{10.92} + \frac{(41-36.26)^2}{36.26} + \frac{(60-48.44)^2}{48.44} + \frac{(20-32.20)^2}{32.20} + \frac{(9-12.10)^2}{12.10} = 8.87$$

4. Región crítica

Para gl = $k - 1 = 5 - 1 = 4$, se tiene: $\chi^2_{crítica}$ (4,0.05) = 9.49
Sí $\chi_o^2 \geq \chi^2_{crítica}$, se rechaza Ho.

5. Conclusión

Los datos están bien descritos por la distribución binomial, con n= 5 y p = 0.4

Ejemplo 107

Se desea confirmar la afirmación de que el porcentaje de cenizas contenidas en carbón producido por cierta empresa, sigue aproximadamente una distribución normal. Los datos presentados a continuación, representan la cantidad porcentual de cenizas encontradas en 250 muestras de carbón analizadas en el laboratorio.

Cenizas (%)	Frecuencia observada
[9.5 – 10.5)	2
[10.5 – 11.5)	5
[11.5 – 12.5)	16
[12.5 – 13.5)	42
[13.5 – 14.5)	69
[14.5 – 15.5)	51
[15.5 – 16.5)	32
[16.5 – 17.5)	23
[17.5 – 18.5)	9
[18.5 – 19.5)	1
	250

¿Cuál es la decisión que se debe tomar con un nivel de 5% de significancia?

$$\hat{\mu} = \bar{x}_{obs} = 14.512$$

$$\hat{\sigma}^2 = s^2{}_{obs} = 2.7007$$

Sea X la variable aleatoria porcentaje de cenizas contenidas en el carbón producido por la empresa, las hipótesis a evaluar son:

Ho : X ~ N (14.512, 2.7007). La variable porcentaje de cenizas sigue una distribución normal, com media igual a 14.512% y desviación estándar de 2.7007%.

Ha: La variable porcentaje de cenizas no sigue una distribución normal.

Procedimiento:

1. Las diversas clases constituyen las k categorías de valores de la variable X y serán numeradas de 1 a 10. Con la finalidad de que la distribución de los datos contenga los valores del intervalo (–∞,∞) correspondientes al modelo normal, se agregan a las categorías 1 y 10 los valores menores de 9.5 y mayores que 19.5, respectivamente.

2. Se calculan las frecuencias esperadas, de la siguiente forma:

$$e_i = n \times p_i \text{ y } p_i = Z = \frac{x_i - \mu}{\sigma}$$

Por ejemplo:

$e_1 = 250 \times p\,(X < 10.5 \,/\, \text{Ho es verdadera})$; $p\,(X < 10.5 \,/\, \text{Ho es verdadera})$

$$= p\left(\frac{x-14.5}{\sqrt{2.7}} < \frac{10.5-14.5}{\sqrt{2.7}}\right) = p(Z < -2.44) = 0.5 - 0.4927 = 0.0073$$

$e_1 = 250 \times 0.0073 = 1.83$

$$e_2 = 250 \times \; p\left(\frac{10.5-14.5}{\sqrt{2.7}} \le \frac{x-14.5}{\sqrt{2.7}} \le \frac{11.5-14.5}{\sqrt{2.7}}\right)$$

$$e_2 = 250 \times \; p\left(-2.44 \le Z \le -1.83\right), \qquad e_2 = 250 \times \left(0.4927-0.4664\right) = 250 \times 0.0263 = 6.58$$

Y así sucesivamente. La tabla final con las frecuencias esperadas por categoría, se presenta a continuación:

Categoría	Cenizas (%)	Frecuencia esperada
1	< 10.5	1.82
2	[10.5 – 11.5)	6.58
3	[11.5 – 12.5)	19.40
4	[12.5 – 13.5)	39.92
5	[13.5 – 14.5)	57.28
6	[14.5 – 15.5)	57.28
7	[15.5 – 16.5)	39.92
8	[16.5 – 17.5)	19.40
9	[17.5 – 18.5)	6.58
10	>18.5	1.82
	Sumatoria	250

3. La aproximación para el modelo χ^2 será mejor si todas las frecuencias esperadas fueren por lo menos iguales a 5. Si esto no sucede para alguna categoría, se debe combinar con otra, de forma conveniente, garantizando que todas las frecuencias esperadas atiendan a ese criterio. De acuerdo con esto, se agrupa la categoría 1 con la 2 y la 9 con la 10. Las nuevas categorías y sus respectivas frecuencias esperadas y observadas se presentan a continuación:

Categoría	Frecuencia esperada	Frecuencia observada
1	8.40	7
2	19.40	16
3	39.92	42
4	57.28	69
5	57.28	51
6	39.92	32
7	19.40	23
8	8.40	10

4. Se efectúa el cálculo de la estadística Ji-cuadrada

$$\chi_o^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \frac{(7-8.40)^2}{8.40} + \frac{(16-19.40)^2}{19.40} + \ldots + \frac{(10-8.40)^2}{8.40} = 6.57$$

Para determinar la región crítica (RC), se utiliza la distribución χ^2 con k – 1– p grados de libertad, siendo k = número de categorías y p el número de parámetros que fueron estimados, para este ejemplo: 8 – 1– 2 = 5 grados de libertad. Con el auxilio de la tabla de χ^2 se obtiene: RC = { χ^2 : $\chi^2 \ge 11.07$ } para $\alpha = 0.05$. Note que RC no contiene χ_o^2 y por lo tanto, se decide por la aceptación del modelo Normal para la variable aleatoria X.

Ejemplo 108

Análisis de datos categóricos, prueba de bondad de ajuste.

Un genetista realiza un cruzamiento de arvejas lisas y amarillas con rugosas y verdes, obteniendo las siguientes frecuencias observadas (Oi):

Número de plantas según el color y la forma de la semilla

Forma de la semilla	Color de la Semilla		Total
	Amarillas	Verdes	
Lisas	1080	210	1290
Rugosas	200	110	310
Total	1280	320	1600

Para saber si estas características siguen una de las leyes clásicas de la herencia mendeliana, se trata de establecer si la frecuencia relativa de cada una de las clases de la población es: 9/16, 3/16, 3/16 y 1/16, respectivamente. Esta misma hipótesis se expresa como "la proporción es 9:3:3:1" (observar que 9+3+3+1=16, por lo que ambas formas son equivalentes). Así:

Ho: la frecuencia es 9:3:3:1, contra
Ha: la frecuencia no es 9:3:3:1

Los valores esperados, si la hipótesis nula es cierta, surgen de multiplicar cada una de las frecuencias relativas (o proporciones) por el total de individuos observados en la muestra. Por lo tanto, la tabla de frecuencias esperadas es:

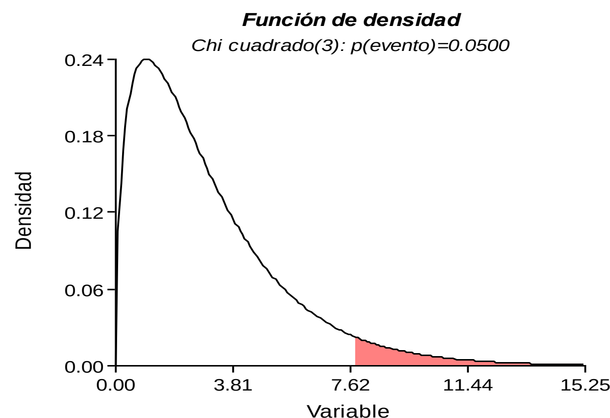
Tipo	Observadas (Oi)	Esperadas (Ei)
Lisas y amarillas	1080	9/16 * 1600 = 900
Lisas y verdes	210	3/16 * 1600 = 300
Rugosas y amarillas	200	3/16 * 1600 = 300
Rugosas y verdes	100	1/16 * 1600 = 100
Total	1600	1600

El estadístico descriptivo como $\chi 2$ tiene una distribución aproximada $\chi 2$ con (r-1-k) grados de libertad, siendo r la cantidad de categorías y k la cantidad de parámetros estimados. En el ejemplo, r = 4 categorías, k=0, luego $\chi 2 \sim \chi 23$. El estadístico evaluado en este caso es:

$$\chi^2 = \sum_{i=1}^{r} \left(\frac{(Oi - Ei)^2}{Ei} \right)$$

$$\chi^2 = \frac{(1080-900)^2}{900} + \frac{(210-300)^2}{300} + \frac{(200-300)^2}{300} + \frac{(110-100)^2}{100} = 97.33$$

La región de rechazo para este contraste está siempre a la derecha, o sea, para valores grandes de χ^2. El cuantil (1-α) de la distribución $\chi 2$ con 3 grados de libertad es 7.81 para α = 0.05. Como 97.33 es mayor que 7.81, se rechaza Ho; por lo tanto, las frecuencias no siguen una distribución 9:3:3:1.



LECTURA

La confusión en medio de la interpretación verdadera; error tipo I (nivel de significancia), error tipo II y el valor p.[1/]

En muchas ocasiones nos encontramos en medio de la confusión al momento de interpretar el valor p, el error tipo I y en ciertos casos el error tipo II; en algunas situaciones se habla del valor p como si se tratara del error tipo I. El presente escrito abordara el concepto de los mencionados anteriormente y se aclarará la diferencia entre valor p y error Tipo I o nivel de significancia.

Para comenzar, describiremos los tipos de errores que se pueden cometer al realizar estadística inferencial (se pretende generalizar los resultados obtenidos en la muestra a la población o universo).

Cuando probamos hipótesis podemos tener alguno de los siguientes resultados:

a) Aceptar una hipótesis verdadera, en este caso estamos en la decisión correcta.

b) Rechazar una hipótesis falsa, estamos en la decisión correcta.

c) Rechazar una hipótesis verdadera, es el error conocido como Tipo I o alfa. Esto equivale a la probabilidad de un resultado erróneo.

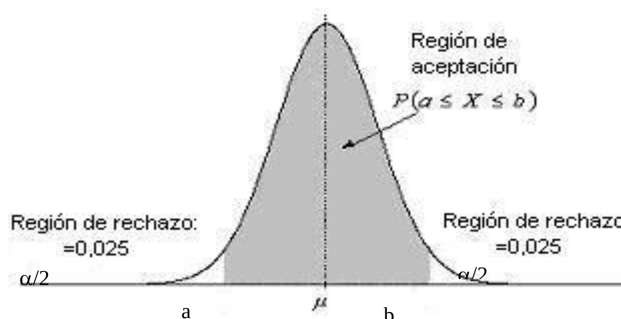


Figura 21. Región de rechazo y aceptación en la distribución normal.

En la Figura 21 se observa que en cada lado de la curva de distribución normal, hay dos pequeñas colas, las cuales son definidas como región de rechazo; es, en esta región donde se acepta la hipótesis alterna (Hipótesis de trabajo) y se rechaza la hipótesis nula (Hipótesis nula: se plantea en términos de igualdad y es la hipótesis que deseamos rechazar. Ej. Efectividad A = Efectividad B), a la región de rechazo también se le conoce como región crítica. Ahora si un investigador informa que sus resultados son estadísticamente significativos, quiere decir que, según la prueba estadística, sus hallazgos podrían ser válidos y replicables con nuevas muestras de sujetos.

En el caso de probar hipótesis, el error tipo I o alfa es establecido por el investigador antes de realizar el proceso de prueba de hipótesis inferenciales. Los valores más comunes de significancia son de 0.05, 0.01, 0.001, estos valores dependen de la rigurosidad que establezca el investigador para su análisis.

Ahora, para determinar el concepto de valor p, iniciemos con la definición clásica; es el valor de probabilidad o "significancia" de los resultados. El valor p mide la probabilidad de obtener un valor para el estadístico tan extremo como el realmente observado si la hipótesis nula fuera cierta.

Con lo anterior podemos ilustrar de manera clara que es realmente el valor p; supongamos que la diferencia observada en la evaluación de efectividad de dos fármacos (tradicional y uno nuevo) es de 15 por ciento a favor del nuevo. Un valor p de 0.02 indicará que, si el nuevo fármaco no ha tenido un verdadero efecto, habría solamente una oportunidad del 2% de obtener una diferencia de 15% o mayor.

Ahora teniendo claro la interpretación del valor p, demos su definición, de la siguiente manera: es la probabilidad asociada a un estadístico de prueba calculado a partir de los datos obtenidos en una investigación, e indica la probabilidad de obtener un valor tan extremo como el estadístico de prueba calculado en cualquier dirección, cuando la hipótesis nula es verdadera. Significa que existe una probabilidad menor que alfa (error tipo I) de que el resultado obtenido sea atribuible al azar, o una certeza del (1−alfa) de que el resultado obtenido por la intervención sea verdadero.

De acuerdo a la definición de valor p, nos queda claro que este valor no es igual al valor alfa o error tipo I, la primera diferencia se observa al momento en que el investigador fija la zona de rechazo o el nivel de significancia alfa, mientras que el valor p viene dado por el estadístico de prueba calculado a partir de los datos de la investigación y puede ser superior, inferior o igual al valor alfa y no es controlado por el investigador, ya que, es un valor asociado al estadístico de prueba.

El valor p para una prueba puede definirse también como el valor más pequeño del error tipo I o alfa por el cual la hipótesis nula se puede rechazar. Si el valor p tiende a ser pequeño, menos fuerza tendrá la hipótesis nula como una explicación de los datos observados. Además el nivel alfa para algunos autores es definido como un nivel de la probabilidad de equivocarse y se fija antes de probar hipótesis inferenciales, es un valor de certeza respecto a no equivocarse. Así, el nivel de significancia representa áreas de riesgo o confianza en la distribución muestral.

En lugar de decir que un valor observado de la estadística de prueba es significativo o no significativo a un valor alfa, muchos autores en sus obras de investigación prefieren informar la probabilidad exacta de obtener un valor tanto o más extremo que el observado, si la hipótesis nula es verdadera. En este momento los autores darán el valor calculado de la estadística de prueba junto con el valor p asociado a esta.

Cabe recordar o dar a conocer para quienes no conocen muy bien del tema, que el nivel de significancia establece el límite de la región de rechazo, por tanto la hipótesis nula en un estudio se rechaza cuando el valor p asociado a la prueba estadística utilizada para contrastar la hipótesis, es inferior al valor alfa establecido por el investigador (valor p < nivel de significancia).

Para terminar, se recomienda utilizar el valor p al momento de referirnos a la significancia estadística, cuando empleamos frases como: "existe diferencia significativa", "existe relación significativa", "es

significativamente diferente de cero". Conclusiones de este tipo deben ir acompañadas del valor p asociado a la prueba, más que del nivel de significancia establecido por el investigador. El valor p como parte de los resultados de una investigación proporciona más información al lector que afirmaciones del tipo: "la hipótesis nula se rechaza en el nivel 0.05 de significancia", "los resultados no son significativos a un nivel de 0.05".

Mientras que el informar el valor p asociado a una prueba permite al lector saber con exactitud que tan probable o no es el valor calculado de la prueba estadística realizada dado que la hipótesis nula es verdadera.

Sugerencias y suposiciones

Los valores p y las computadoras han suprimido la necesidad de buscar valores en las tablas de la distribución z o t, y eliminan el trabajo tedioso de las pruebas de hipótesis. Advertencia: cuanto más pequeño sea el valor p, mayor será la significancia del estudio. Sugerencia: se puede evitar la confusión aquí al recordar que un valor p es la probabilidad de que el resultado obtenido haya podido ocurrir por el error de muestreo: así, los valores p más pequeños significan menor posiblidad de error de muestreo y mayor significancia.

[1/] Fuente: Héctor Fabio Mueses M. Colegio Odontológico Colombiano, Sede Cali. Marzo de 2003.

LISTA DE EJERCICIOS 9

1. La vida media de una muestra de 100 focos de cierta marca es 1,615 horas. Por similitud con otros procesos de fabricación, se supone que la desviación estándar es igual a 120 horas. Utilizando α=0.05, se desea evaluar si la duración media de todos los focos de esa marca es igual o es diferente a 1,600 horas. ¿Cuál es la conclusión?

2. Un ingeniero forestal desea comparar la dominancia de dos especies nativas. En un levantamiento con 31 parcelas, la especie A presentó dominancia media de 5.3 m^2/ha (con desviación estándar de 1.2) y la especie B presentó un valor medio de 6.7 m^2/ha (y desviación estándar de 2.1). Establezca las hipótesis estadísticas apropiadas, evalúelas y emita sus conclusiones.

3. Los siguientes datos corresponden a los pesos en kilogramos de 15 trabajadores escogidos al azar: 72, 68, 63, 75, 84, 91, 66, 75, 86, 90, 62, 87, 77, 70, 69. Pruebe la H_o: $\mu \geq 74$ con un nivel de significancia de 0.05.

4. Se obtiene una muestra de 16 estudiantes con una $\bar{x}$ = 68 puntos y una varianza de s^2 = 9, en la evaluación final de Estadística I. Hay evidencia suficiente que apoye que la media poblacional de las calificaciones de estadística es mayor de 70 con α = 0.01

5. La producción diaria de una planta industrial de fertilizantes químicos en Teculután (Zacapa), registrada durante n=50 días tiene una media muestral de 871 toneladas y una desviación estándar de 21 toneladas. Pruebe la hipótesis de que el promedio de la producción diaria del producto químico es de μ=880 toneladas, contra la alterna de que μ es mayor o menor que 880 toneladas.

6. La vida media de una muestra de 100 focos marca ANTILLON es 1,615 horas. Por similitud con otros procesos de fabricación, se supone que la desviación estándar es igual a 120 horas. Utilizando α=0.05, se desea evaluar si la duración media de todos los focos de esa marca es igual o es diferente a 1,600 horas. ¿Cuál es la conclusión?

7. Se encuentra que la concentración promedio de zinc que se saca del agua a partir de una muestra de mediciones de zinc en 36 sitios diferentes es de 2.6 gramos por mililitro. Suponga que la desviación estándar de la población es 0.3 gramos por mililitro. ¿Existe suficiente evidencia estadística para decir que la concentración promedio de zinc es menor de 2.9 gramos por mililitro? Utilice α =0.02.

8. Para tratar de estimar la media de consumo por cliente, en el restaurante “Chicken Grill”, se reunieron datos de una muestra de 50 clientes durante un periodo de tres semanas. Si la media de la muestra es de $ 22.60 dólares, con una desviación estándar de $ 7.00. ¿Existe evidencia para decir que la media de la población es mayor de 25 dólares? Pruebe con α = 0.05.

9. El enorme crecimiento de la industria de la langosta en los últimos 20 años, la ha colocado en el segundo lugar de la industria pesquera del estado de la Florida. Hace algunos años se supuso que una declaración por parte del gobierno de las Bahamas que prohibía a los pescadores de langostas de Estados Unidos operar en la parte de la plataforma continental perteneciente a ese país, reduciría notablemente la cantidad de langostas (en libras) obtenida por trampa.

   Según los registros, la captura promedio por trampa es de 30.31 libras. Una muestra aleatoria de 20 trampas para langosta, fue colocada desde que la restricción por parte de las Bahamas entró en vigor, dio los siguientes resultados (pesos expresados en libras):

17.4	18.9	39.6	34.4	19.6
33.7	37.2	43.4	41.7	27.5
24.1	39.6	12.2	25.5	22.1
29.3	21.1	23.8	43.2	24.4

¿Proporcionan estos datos evidencia suficiente que apoye la opinión de que las capturas medias por trampa disminuyeron después de la imposición de las restricciones por parte de las Bahamas? Haga la prueba utilizando un nivel de significancia del 1%

10. MicroPCSystems estimó el año pasado que el 35% de los compradores potenciales de software planeaba esperar hasta que se liberara una actualización de Windows Planet para comprar el nuevo sistema operativo. Después de una campaña publicitaria para dar confianza al público, MicroPCSystems encuestó a 3,000 personas y encontró que 950 todavía se mostraban renuentes. Con el 5% de nivel de significancia, ¿puede la compañía concluir que la proporción de personas renuentes ha disminuido?

11. Un supervisor de control de calidad en una enlatadora sabe que la cantidad exacta en cada lata varía, pues hay ciertos factores imposibles de controlar que afectan a la cantidad de llenado. El llenado medio por lata es importante, pero igualmente importante es la variación σ^2 de la cantidad de llenado. Si σ^2 es grande, algunas latas contendrán muy poco, y otras, demasiado. A fin de estimar la variación del llenado en la enlatadora, el supervisor escoge al azar 10 latas y pesa el contenido de cada una, obteniendo los siguientes pesos (en onzas):

7.96 7.90 7.98 8.01 7.97 7.96 8.03 8.02 8.04 8.02

Suponga que las agencias reguladoras especifican que la desviación estándar de la cantidad de llenado debe ser menor que 0.1 onzas. ¿Esta información proporciona pruebas suficientes de que la desviación estándar de las mediciones de llenado es menor que 0.1 onzas si el nivel de significación queda fijado en un 5%?

12. Se obtiene una muestra de n = 10 tabletas de aspirinas. Cada una de las tabletas fue pesada, obteniendo los siguientes pesos expresados en gramos: 1.19; 1.23; 1.18; 1.21; 1.27; 1.17; 1.15; 1.14; 1.19; 1.2. Los datos proporcionados por la muestra apoyan la afirmación que σ^2 =0.0015 con un nivel de significación fijado en un 1%?.

13. Se seleccionó una muestra aleatoria de n=22 observaciones de una población normal. La media y la varianza muestral eran: 4.13 y 14.14, respectivamente. ¿Es esta evidencia suficiente para indicar que la varianza es menor de 25 u^2? Realice la prueba con un nivel de significancia del 5%.

14. Una muestra aleatoria de n=25 observaciones de una población normal produjo una varianza muestral igual a 21.4 u^2. ¿Proporcionan estos datos evidencia suficiente para indicar que σ^2> 15? Utilice α=0.05

15. Una organización de créditos y seguros agrícolas ha desarrollado un nuevo método de alta tecnología para capacitar al nuevo personal de ventas. El gerente general de la compañía obtuvo una muestra de 16 empleados capacitados de la manera original y encontró ventas diarias promedio de $688 con desviación estándar de la muestra de $32.63. También tomó una muestra de 11 empleados capacitados con el método nuevo y encontraron un promedio de ventas diarias de $706 con desviación estándar de la muestra de $24.84. Para α=0.05 ¿puede la compañía concluir que el promedio diario de ventas aumenta con el nuevo plan?

16. Un investigador desea verificar si la altura de un árbol en pie, medida usando el método trigonométrico (aproximado) no difiere de la altura de un árbol medida en el suelo. Con ese objetivo, midió la altura de 12 árboles por el método trigonométrico, luego los derrumbó y midió nuevamente sus alturas, obteniendo los resultados que se presentan en la siguiente tabla.

No. Árbol	Árbol en pie	Árbol en el suelo
1	20.4	21.7
2	25.4	26.3
3	25.6	26.8
4	26.6	26.2
5	28.6	27.3
6	28.7	29.5
7	29.0	32.0
8	29.8	30.9
9	30.5	32.3
10	30.9	32.3
11	31.1	31.7
12	25.6	28.1

Utilizando un nivel de significancia del 5%, verifique la hipótesis del investigador.

17. Para evaluar el nivel de tensión ocasionada por exámenes escolares, doce alumnos fueron seleccionados y su pulsación medida antes y después del examen. Los datos obtenidos fueron los siguientes:

Instante de la medición	Estudiante											
	1	2	3	4	5	6	7	8	9	10	11	12
Antes	87	78	85	93	76	80	82	77	91	74	76	79
Después	83	84	79	88	75	81	74	71	78	73	76	71

Realice una prueba, con un nivel de significancia del 1% para verificar si existe mayor tensión (o sea, mayor pulsación) antes de la realización de los exámenes. Indique las suposiciones necesarias.

18. Muestras aleatorias e independientes de dos poblaciones normales presentaron las siguientes varianzas:

Población	Tamaño de la muestra	Varianza muestral
1	16	55.7
2	20	31.4

¿Proporcionan los datos evidencia suficiente para indicar que σ_1^2 difiere de σ_2^2 ? Utilice α=0.05

19. Una empresa fabrica y distribuye tres tipos de cerveza: ligera, clara y obscura. En un análisis de segmentación de mercado para las tres cervezas, el grupo de investigación ha planteado la duda de si las preferencias para los tres tipos de cerveza son diferentes entre los consumidores hombres y mujeres. Los resultados de la encuesta realizada, se presentan a continuación:

Sexo	Cerveza preferida			
	Ligera	Clara	Obscura	
Masculino	20	40	20	80
Femenino	30	30	10	70
	50	70	30	150

Con esta información evalúe las siguientes hipótesis:

Ho: La preferencia de cerveza es independiente del sexo del consumidor.
Ha: La preferencia de cerveza no es independiente del sexo del consumidor.

20. En las arvejas, el cotiledón de color amarillo es dominante sobre el verde y la vaina gruesa es dominante sobre la vaina delgada. Cuando ambos caracteres fueron considerados conjuntamente en dihíbridos autofecundado, apareció en la progenie en la siguiente proporción:

	Gruesa	Delgada	Total
Amarillo	556	184	740
Verde	193	61	254
Total	749	245	994

Compruebe si los datos siguen la distribución esperada para la F1, 9 : 3 : 3 : 1 para: amarillo, vaina gruesa; amarillo, vaina delgada; verde, vaina gruesa; verde, vaina delgada, respectivamente.

21. Compare el DAP promedio de los árboles de Teca (*Tectona grandis*) de dos parcelas permanentes de muestreo, ubicadas en los proyectos: Sepila y Viejo, de la finca Seshan, Chahal, Alta Verapaz. La medición fue realizada en el año 2005, en árboles de 7 años.

Proyecto Sepila			Proyecto Viejo		
16.3	19	19.2	23	22	18.4
15	14.8	19.5	16	18.3	22
18.8	15.5	13.5	20	27	22.1
13.4	15.4	19.2	22	22.4	25.5
16.6	15.5	15.6	25	18.5	21.7
14.5	17.5	18.5	22.4	16.6	
17.2	17	19.7	22.6	21.9	

22. El diario “La Nación” de Costa Rica realizó una encuesta, considerando una muestra de 1545 hombres y 1691 mujeres para comparar la cantidad de labores domésticas hechas por mujeres y por hombres en matrimonios con doble aportación económica. El estudio indicó que el 67.5% de los hombres sentían que la división de tareas domésticas era justa, y que el 60.8% de las mujeres sentían que la división era justa.

Con esta información, ¿se puede afirmar con un 95% de confianza, que es mayor la proporción de hombres que sentían que la división del trabajo doméstico era justa, que la proporción correspondiente de mujeres?

23. En la finca Seshan, ubicada en Chahal, Alta Verapaz, se realizaron mediciones de diámetro a la atura del pecho (DAP) y altura, a los árboles de teca (*Tectona grandis*) que se encuentran en las parcelas permanentes de muestreo (PPM). Las mediciones se realizaron en el año 2005 y 2006, cuando los árboles tenían 5 y 6 años, respectivamente. Los datos que se presentan a continuación, corresponden a la PPM 1 ubicada en el Proyecto Sepila:

No.	Árbol correlativo	DAP (cm)		ALTURA (m)	
		2005	2006	2005	2006
1	4	20.1	22.5	19.5	20.0
2	6	17.1	18.5	18.0	21.5
3	12	19.5	21.0	18.7	21.0
4	13	20.0	21.5	19.0	21.0
5	16	25.0	25.6	19.5	20.5
6	22	20.8	23.5	19.0	20.0
7	23	21.6	22.6	19.0	19.5
8	25	21.0	23.1	20.3	21.5
9	27	20.0	21.5	19.0	21.0
10	28	19.6	22.0	20.0	21.0
11	32	17.4	18.8	19.5	21.0
12	34	19.5	20.5	19.3	20.0
13	35	16.9	18.9	20.5	21.0
14	30	16.5	17.9	22.0	23.0
15	40	19.4	21.0	21.0	22.0
16	43	20.9	22.5	18.3	19.0
17	46	21.9	23.0	17.5	22.5
18	48	23.5	27.0	21.3	22.5
19	51	19.4	20.7	20.0	21.5

a) Compare si el incremento de DAP fue significativo de un año a otro
b) Compara si el incremento de altura fue significativo de un año a otro.

En ambos casos utilice un nivel de 5% de significancia. Redacte sus conclusiones.

24. A continuación se presentan las notas obtenidas por alumnos de dos secciones del laboratorio del curso de Estadística, en el primer parcial.

Sección A	
70	90
90	90
100	100
90	45
20	90
80	72
45	50
80	30

Sección B		
80	100	50
80	70	35
50	60	60
20	90	70
90	50	90
75	50	40
90	70	80
80	50	70

De acuerdo con los resultados obtenidos, ¿se puede afirmar que no existe diferencia entre las notas promedio de los dos grupos? Utilice un nivel de 5% de significancia.

UNIDAD VI
ANÁLISIS DE CORRELACION LINEAL SIMPLE

6.1 INTRODUCCION

Frecuentemente estamos interesados en estudiar la manera como dos variables están asociadas y cuantificar ese grado de asociación. Por ejemplo:

- ¿Será que plantas con la parte aérea más desarrollada tienden a tener el sistema radicular más desarrollado?
- ¿Será que la materia seca de la parte aérea de la planta de okra está relacionada con la materia seca de las raíces? O aún, ¿será que esas dos variables crecen en el mismo sentido?
- ¿Será que el contenido de azúcar en plantas de caña está asociado con el contenido de humedad en el suelo?
- ¿Será que las variables: largo del cuerpo y profundidad del tórax en vacas lecheras están asociadas?

Para responder a cuestiones de esta naturaleza, se utilizan las siguientes medidas: Covarianza y el coeficiente de correlación momento-producto de Pearson.

6.2 COVARIANZA

El estimador de la covarianza para una muestra de *n* pares de observaciones, es dado por:

$$C\hat{o}v\,[X,Y] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i\, y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right]$$

La covarianza ofrece una idea del signo y del grado de intensidad de la relación entre dos variables, a través de su variación conjunta. Sin embargo, ella puede asumir cualquier valor real, lo que dificulta su interpretación. Para solucionar este inconveniente, se utiliza una medida más efectiva, el COEFICIENTE DE CORRELACION DE PEARSON.

6.3 COEFICIENTE DE CORRELACION DE PEARSON

Definición

Sean X y Y dos variables aleatorias con varianzas σ_x^2 y σ_y^2 Cov (X,Y), se define el coeficiente de correlación de Pearson entre X y Y por:

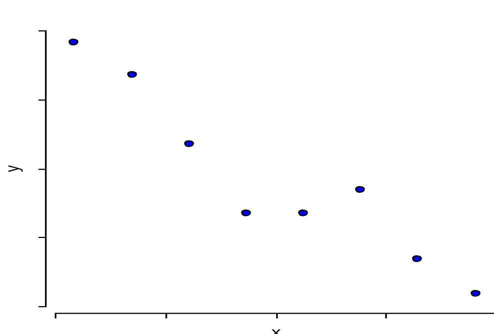
$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\sigma^2_X \times \sigma^2_Y}}, \; -1 \le \rho \le 1.$$

Cuyo estimador es dado por la siguiente expresión: $\hat{\rho} = \frac{\hat{Cov}(X,Y)}{\sqrt{S^2_X \times S^2_Y}}$

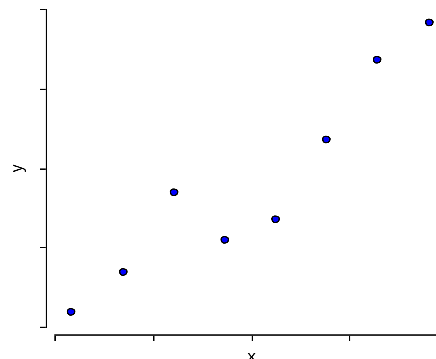
Equivalente a:

$$r = \frac{\sum_{i=1}^{n} x_i\, y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\ \sqrt{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}} \quad , -1 \le r \le 1.$$

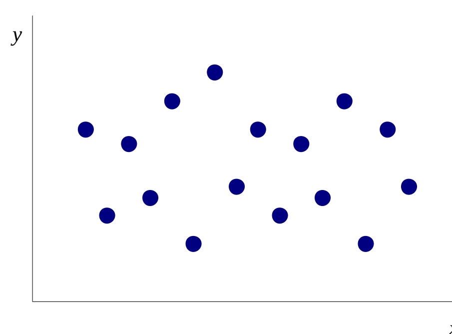
El coeficiente de correlación de Pearson presenta la ventaja de ser un valor entre –1 y 1, facilitando su interpretación. Esta asociación será tan grande, cuando r esté más alejado del valor cero (0) y nula cuando $r = 0$. A continuación se presentan ejemplos de diagramas de dispersión y correspondientes coeficientes de correlación lineal de Pearson.



(a) Fuerte correlación lineal negativa, r está cerca de –1.



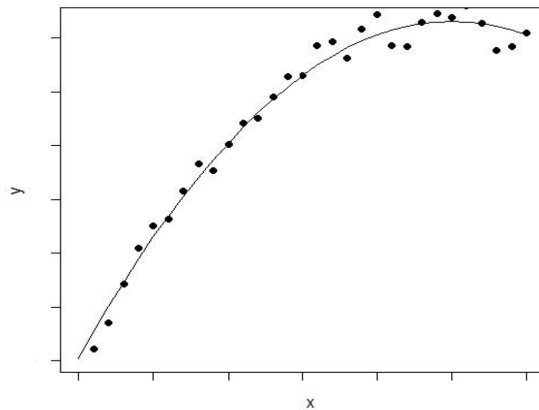
(b) Fuerte correlación lineal positiva, r está cerca de 1.



(c) Ninguna correlación lineal aparente, r está cerca de cero.

🗎 IMPORTANTE:

El coeficiente de correlación lineal es un indicador del grado de asociación lineal entre dos variables. Así, $\rho = 0$ no indica ausencia de asociación entre las variables, pero sí, ausencia de asociación lineal entre las variables. A continuación se presentan algunos diagramas de dispersión en que $r = 0$ y existe relación entre las variables.



(d) Correlación no lineal

6.4 INFERENCIA ACERCA DEL COEFICIENTE DE CORRELACION

A continuación se detallan los pasos necesarios para la realización de la prueba de hipótesis para verificar si el coeficiente de correlación es estadísticamente diferente de cero.

1. Hipótesis a ser evaluadas.

   Ho: $\rho = 0$ (No hay correlación lineal)
   Ha: $\rho \neq 0$

2. Estadística de la prueba

Suponiendo que la muestra fue extraída de una población con distribución normal bivariada, la estadística:

$$t_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \ ,$$

que tiene distribución *t* de student con *n*-2 grados de libertad, donde *n* es el tamaño de la muestra y *r* el coeficiente de correlación muestral, puede utilizarse para probar la hipótesis nula Ho: $\rho = 0$.

3. Dado el nivel de significancia α, construir la región critica de la prueba.

$$t\left(n-2, \frac{\alpha}{2}\right)$$

4. Regla de decisión: Rechazar Ho sí $|t_{obs}| \geq$ *t crítico*

Ejemplo 109

En el valle de Almolonga, Quetzaltenango, se realizó un estudio para medir los contenidos de algunos elementos químicos en el suelo. Se tomó como unidad de muestreo, una parcela de 441 m^2, y se extrajeron dos muestras compuestas de suelo, para dos profundidades: 0 a 15 cm y 15 a 30 cm.
A continuación se presentan los resultados de un muestreo de suelos para la profundidad: 0 a 15 cm, y para los contenidos de materia orgánica (en porcentaje) y calcio (meq/100 gramos), obtenidos en 36 muestras tomadas al azar:

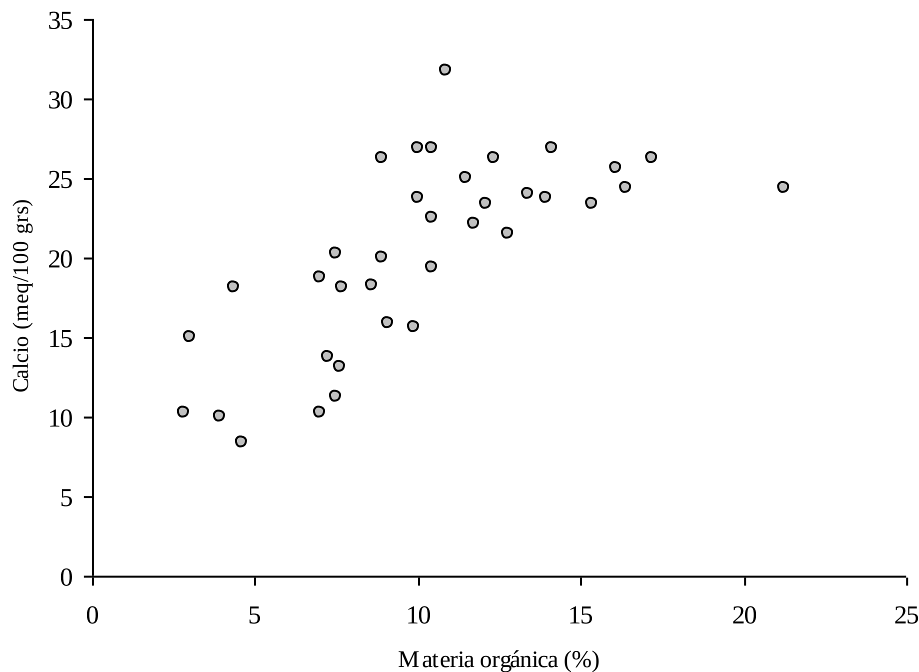
Muestra	MO (x)	Ca (y)	x^2	y^2	xy
1	2.82	10.28	7.95	105.68	28.99
2	3.04	14.96	9.24	223.80	45.48
3	3.91	9.98	15.29	99.60	39.02
4	4.39	18.09	19.27	327.25	79.42
5	4.61	8.42	21.25	70.90	38.82
6	7.03	10.30	49.42	106.09	72.41
7	7.03	18.72	49.42	350.44	131.60
8	7.25	13.73	52.56	188.51	99.54
9	7.47	11.23	55.80	126.11	83.89
10	7.47	20.28	55.80	411.28	151.49
11	7.61	13.09	57.91	171.35	99.61
12	7.69	18.09	59.14	327.25	139.11
13	8.57	18.20	73.44	331.24	155.97
14	8.91	26.19	79.39	685.92	233.35
15	8.92	19.97	79.57	398.80	178.13
16	9.10	15.91	82.81	253.13	144.78
17	9.88	15.60	97.61	243.36	154.13
18	10.00	23.71	100.00	562.16	237.10
19	10.00	26.82	100.00	719.31	268.20
20	10.43	19.33	108.78	373.65	201.61
21	10.43	22.45	108.78	504.00	234.15
22	10.43	26.82	108.78	719.31	279.73
23	10.87	31.80	118.16	1011.24	345.67
24	11.46	24.96	131.33	623.00	286.04
25	11.74	22.13	137.83	489.74	259.81
26	12.09	23.40	146.17	547.56	282.91
27	12.35	26.20	152.52	686.44	323.57
28	12.78	21.53	163.33	463.54	275.15
29	13.41	24.02	179.83	576.96	322.11
30	13.97	23.71	195.16	562.16	331.23
31	14.14	26.83	199.94	719.85	379.38
32	15.38	23.40	236.54	547.56	359.89
33	16.10	25.58	259.21	654.34	411.84
34	16.41	24.33	269.29	591.95	399.26
35	17.17	26.19	294.81	685.92	449.68
36	21.28	24.32	452.84	591.46	517.53
Suma	366.14	730.57	4329.19	16050.86	8040.60

Solución:

- Cálculo del coeficiente de correlación (r):

$$r = \frac{8040.60 - \frac{(366.14)(730.57)}{36}}{\sqrt{4329.19 - \frac{(366.14)^2}{36}}\sqrt{16050.86 - \frac{(730.57)^2}{36}}} = 0.709$$

El valor de r (0.71) indica que la materia orgánica y el calcio tienen una asociación directa o positiva; esto se observa al construir el diagrama de dispersión:



- Evaluación de la hipótesis Ho: $\rho = 0$ contra Ha: $\rho \neq 0$, utilizando un nivel de significancia del 5%.

$$t_{obs} = \frac{0.709 \times \sqrt{36-2}}{\sqrt{1-0.709^2}} = 5.86$$

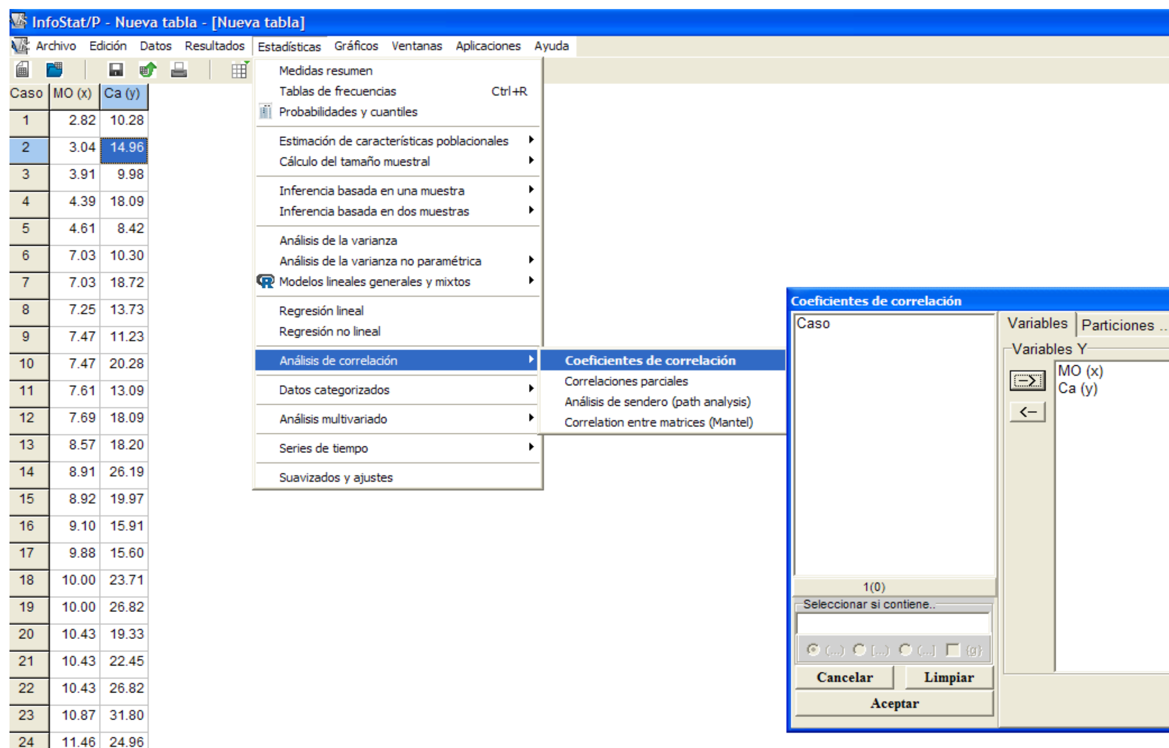
Para calcular el valor crítico de t (34,0.05), puede utilizar la función estadística de MS Excel DISTR.T.INV (0.05,34), como resultado da: 2.03. Como el valor de t observado excede el valor crítico de t = 2.03, se concluye que hay evidencia suficiente para señalar que existe correlación lineal entre el contenido de materia orgánica y el contenido de calcio.

Otra manera de concluir, es por medio del cálculo del valor p (*p value*), denominado también como alfa estimado. Para ello utilice la función estadística de MS Excel DISTR.T (5.86,34,2); el primer valor corresponde al valor de t observado, el segundo al número de grados de libertad, y el

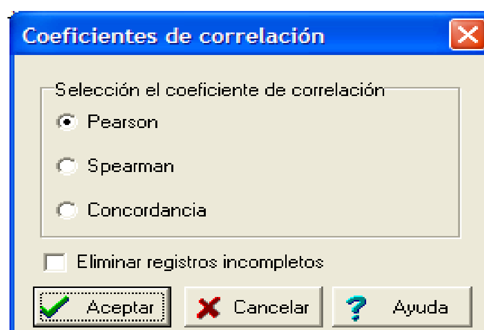
tercero al número de colas, en este caso 2, por ser una prueba bilateral. El valor resultante es: 0.0000013; que es menor al valor de alfa (0.05) definido. Por lo que se llega a la misma conclusión.

Resolución de Ejemplo 109 con Infostat

1. Ingreso de datos



2. Solicitar análisis



3. Verificar resultados y comparar con los cálculos manuales

Coeficientes de correlación

Correlación de Pearson: coeficientes\probabilidades

	MO (x)	Ca (y)
MO (x)	1.00	1.3E-06
Ca (y)	0.71	1.00

Ejemplo 110

Luego de una fuerte tempestad el 1º de febrero de 1898, diversos gorriones moribundos fueron llevados al laboratorio biológico de Hermon Bumpus en la Universidad de Brown en Rhode Island. Subsecuentemente cerca de la mitad de los pájaros murieron, y Bumpus vio eso como una oportunidad de encontrar soporte para la teoría de la selección natural de Charles Darwin. Para ese fin, él hizo ocho medidas morfológicas en cada pájaro, y también los pesó. Los resultados de cinco de las medidas son mostrados en la siguiente tabla, para hembras únicamente:

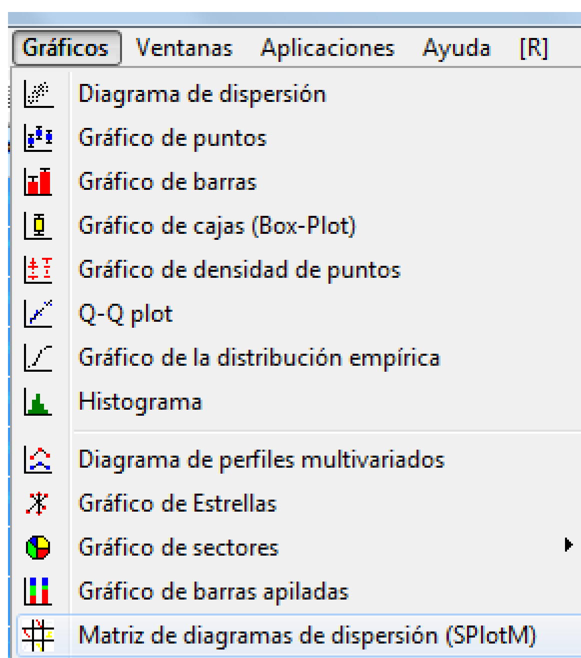
Pájaro	X1	X2	X3	X4	X5
1	156	245	31.6	18.5	20.5
2	154	240	30.4	17.9	19.6
3	153	240	31	18.4	20.6
4	153	236	30.9	17.7	20.2
5	155	243	31.5	18.6	20.3
6	163	247	32	19	20.9
7	157	238	30.9	18.4	20.2
8	155	239	32.8	18.6	21.2
9	164	248	32.7	19.1	21.1
10	158	238	31	18.8	22
11	158	240	31.3	18.6	22
12	160	244	31.1	18.6	20.5
13	161	246	32.3	19.3	21.8
14	157	245	32	19.1	20
15	157	235	31.5	18.1	19.8
16	156	237	30.9	18	20.3
17	158	244	31.4	18.5	21.6
18	153	238	30.5	18.2	20.9
19	155	236	30.3	18.5	20.1
20	163	246	32.5	18.6	21.9
21	159	236	31.5	18	21.5
22	155	240	31.4	18	20.7
23	156	240	31.5	18.2	20.6
24	160	242	32.6	18.8	21.7
25	152	232	30.3	17.2	19.8
26	160	250	31.7	18.8	22.5
27	155	237	31	18.5	20
28	157	245	32.2	19.5	21.4
29	165	245	33.1	19.8	22.7
30	153	231	30.1	17.3	19.8
31	162	239	30.3	18	23.1
32	162	243	31.6	18.8	21.3
33	159	245	31.8	18.5	21.7
34	159	247	30.9	18.1	19

35	155	243	30.9	18.5	21.3
36	162	252	31.9	19.1	22.2
37	152	230	30.4	17.3	18.6
38	159	242	30.8	18.2	20.5
39	155	238	31.2	17.9	19.3
40	163	249	33.4	19.5	22.8
41	163	242	31	18.1	20.7
42	156	237	31.7	18.2	20.3
43	159	238	31.5	18.4	20.3
44	161	245	32.1	19.1	20.8
45	155	235	30.7	17.7	19.6
46	162	247	31.9	19.1	20.4
47	153	237	30.6	18.6	20.4
48	162	245	32.5	18.5	21.1
49	164	248	32.3	18.8	20.9

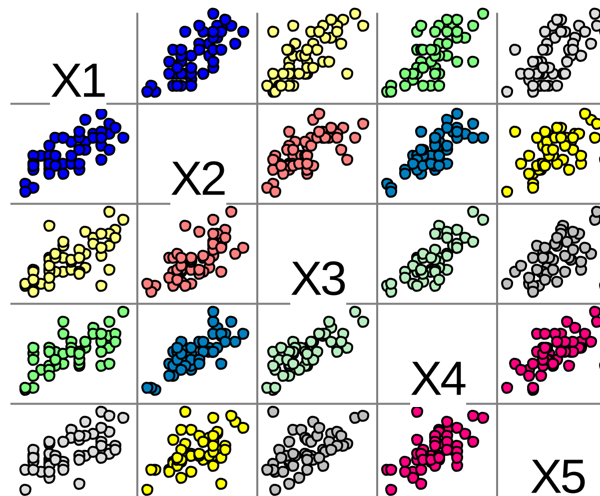
Nota:
X1 = largo total (mm)
X2 = extensión alar (mm)
X3 = largo del pico y la cabeza (mm)
X4 = largo del húmero
X5 = largo de la quilla del esternón

Fuente: Adaptado de Bumpus, H.C. (1898), Biological Lectures, 11th Lecture, Marine Biology Laboratory, Woods Hole, MA, pp. 209-226.

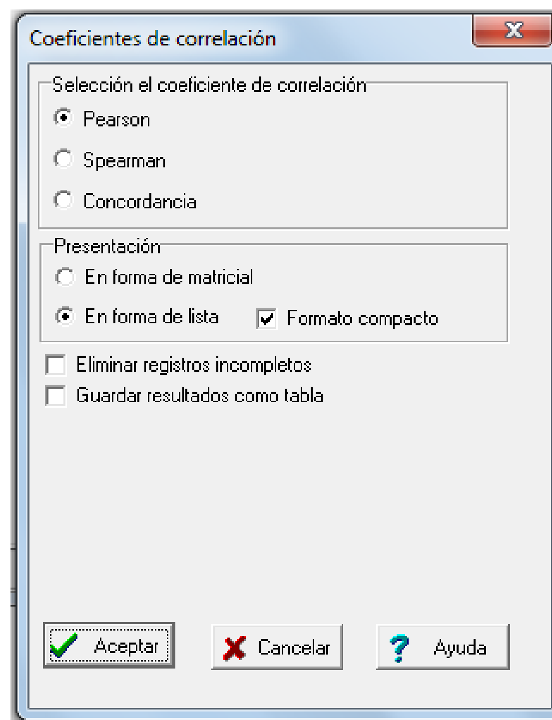
a) Construya en Infostat los gráficos de dispersión para ver las posibles asociaciones entre las variables.



Matriz de diagramas de dispersión



b) Calcule los valores del coeficiente de correlación de Pearson y su significancia



Coeficientes de correlación

Correlación de Pearson

Variable(1)	Variable(2)	n	Pearson	p-valor
X1	X2	49	0.73	<0.0001
X1	X3	49	0.66	<0.0001
X1	X4	49	0.65	<0.0001
X1	X5	49	0.61	<0.0001
X2	X3	49	0.67	<0.0001
X2	X4	49	0.77	<0.0001
X2	X5	49	0.53	0.0001
X3	X4	49	0.76	<0.0001
X3	X5	49	0.53	0.0001
X4	X5	49	0.61	<0.0001

c) Conclusiones:

Todas las variables presentan una asociación lineal directa significativa, ya que los valores de p son menores que el nivel de significancia (0.05). Esto se pudo constatar también gráficamente. Las correlaciones más fuertes se dan entre X2 (extensión alar) y X4 (largo del húmero), con r=0.77, y X3 (largo del pico y cabeza) y X4 (largo del húmero), con r=0.76. Por otra parte, los menores valores de asociación se da entre X2 (extensión alar) y X5 con r=0.53, y X3 (largo del pico y cabeza) y X5 (largo de la quilla del esternón) con r=0.53.

LISTA DE EJERCICIOS 10

1. Referente al estudio realizado en el valle de Almolonga, Quetzaltenango, para medir los contenidos de algunos elementos químicos en el suelo. A continuación se presentan los resultados para la profundidad de 15 a 30 centímetros. Realice las correlaciones, dos a dos, entre los elementos que se presentan. Construya los diagramas de dispersión y evalúe la significancia estadística, utilizando el valor crítico de t y el valor p.

% M.O.	PH	mg.kg-1 P	mg.kg-1 K	meq/100g Ca	meq/100g Mg	mg.kg-1 Fe	mg.kg-1 Cu	mg.kg-1 Zn	mg.kg-1 Mn
7.7	6.0	547	215	24.32	5.29	28.5	0.5	150	30.5
10.09	5.9	485	188	24	5.08	23	0.5	195	36.5
10.31	7.2	565	150	28.69	6.93	12.5	0.1	63	31.5
9.4	7.0	525	220	25.57	5.08	22	0.5	125	56.5
5.35	6.9	401	270	14.03	3.28	65.5	2.5	70	45.5
14.54	6.9	376	103	24.96	9.46	9.5	0.1	30	35.0
11.26	6.6	569	128	24.96	9	10.5	0.1	140	30.5
12.35	6.0	498	363	22.46	5.76	24.5	0.5	112	67.5
3.06	6.8	183	303	7.48	2.31	107.5	2.5	35	57.0
9.48	6.6	510	725	20.9	5.65	28.5	0.5	185	44.0
13.84	6.0	482	188	21.84	4.27	23.5	0.5	97.5	45.5
15.45	6.0	391	218	12.48	6.84	8.5	0.5	118	31.5
10.84	5.6	511	325	23.09	4.27	24.5	1	160	47.0
8.49	6.6	478	140	22.46	7.14	20.5	0.5	112	53.5
13.13	6.4	479	120	23.71	5.81	10.5	0.5	150	51.5
11.26	6.5	482	140	24.34	8.74	8.5	0.5	75	35.5
8.28	6.8	318	128	22.15	9.15	10	0.1	145	43.5
11.99	5.3	446	205	18.72	3.8	32	0.5	205	64.0
6.33	5.0	523	233	9.67	1.85	73	1.5	115	44.0
8.57	5.1	362	120	16.54	4.78	24.5	0.5	160	39.5
8.49	6.8	661	343	28.06	5.6	20.5	0.1	103	33.0
15.66	5.9	459	358	29.31	6.32	6.5	0.1	148	34.0
9.66	5.9	284	95	19.01	4	28	0.5	138	46.5
2.87	6.9	254	153	12.78	3.08	87.5	2.5	27	52.0
10.57	5.7	483	360	22.13	5.19	22	0.5	150	47.5
6.79	5.8	345	208	14.03	3.03	43	1	83	43.5
9.49	6.3	179	235	18.09	3.19	28.5	1.5	118	165.0
8.83	6.9	328	130	21.84	4.93	25.5	0.5	108	112.5
6.19	6.3	528	305	19.03	3.86	36	1	108	50.5
6.88	5.9	394	250	15.91	4.11	40.5	0.5	72.5	43.5
7.51	5.3	303	303	16.84	4.57	39	1.5	180	51.0
3.69	5.3	350	350	7.18	2.42	60.5	2	38	33.0
5.54	5.3	381	148	9.05	2.21	36.5	1	73	35.0
4.22	6.7	407	283	14.35	2.16	23.5	0.5	63	33.0
8.04	5.0	536	333	14.04	3.29	54	1.5	190	74.0
8.61	5.3	537	400	19.03	5.19	45.5	0.5	90	43.0

Nota: Puede ser que algunas correlaciones no sean significativas. Investigue la interpretación práctica de estas correlaciones.

2. A continuación son presentados los pesos de materia seca de la raíz y de la parte aérea (hojas y tallos) de 56 plantas de palmito (*Euterpe edullis* Mart.) provenientes de semillas de plantas del Parque Estatal Carlos Botelho (Floresta Ombrófila Densa o Atlantic Rainforest), localizado en el Estado de Sao Paulo, Brasil.

Planta	Raiz	Parte aérea	Planta	Raiz	Parte aérea
1	0.091	0.211	29	0.162	0.305
2	0.067	0.247	30	0.237	0.513
3	0.086	0.333	31	0.104	0.103
4	0.130	0.396	32	0.132	0.427
5	0.196	0.465	33	0.173	0.375
6	0.091	0.248	34	0.132	0.427
7	0.146	0.332	35	0.173	0.375
8	0.122	0.375	36	0.103	0.353
9	0.168	0.412	37	0.112	0.320
10	0.101	0.305	38	0.059	0.142
11	0.194	0.565	39	0.109	0.314
12	0.265	0.493	40	0.103	0.211
13	0.175	0.391	41	0.141	0.291
14	0.230	0.672	42	0.119	0.292
15	0.230	0.672	43	0.092	0.383
16	0.111	0.355	44	0.072	0.071
17	0.114	0.252	45	0.126	0.209
18	0.092	0.298	46	0.117	0.463
19	0.103	0.32	47	0.153	0.339
20	0.145	0.327	48	0.153	0.421
21	0.250	0.763	49	0.267	0.641
22	0.220	0.42	50	0.126	0.309
23	0.211	0.542	51	0.099	0.285
24	0.121	0.235	52	0.067	0.216
25	0.051	0.123	53	0.314	0.867
26	0.087	0.28	54	0.139	0.275
27	0.148	0.255	55	0.314	0.867
28	0.112	0.166	56	0.139	0.275

Verifique si existe correlación lineal simple entre las dos variables estudiadas.

3. Se cuenta con los registros de diámetro a la altura del pecho (en cms) y la altura (en m) de 10 árboles de *Eucalyptus grandis* plantados en un rodal de Rivera, Uruguay.

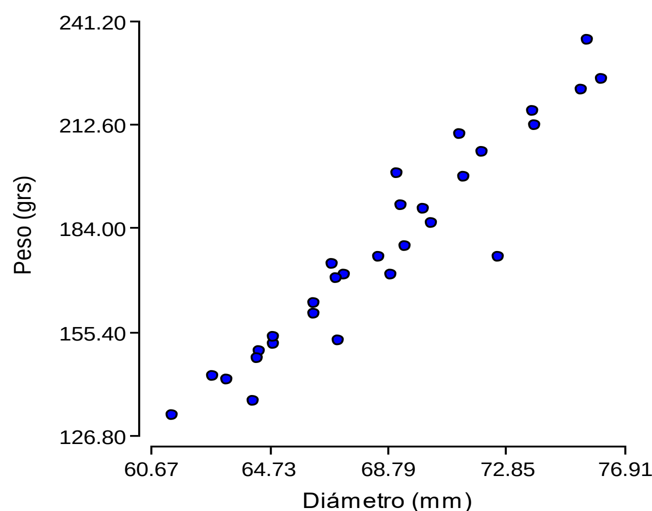
DAP	22.85	22.59	17.42	19.81	19.91	18.69	21.86	21.23	19.17	19.39
Altura	22.45	20.54	15.65	17.97	17.74	18.61	21.21	22.44	18.89	19.85

a) Construya un diagrama de dispersión.
b) Calcule la covarianza muestral.
c) Calcule el coeficiente de correlación lineal de Pearson en la muestra.
d) Realice una prueba de hipótesis para probar que el coeficiente de correlación lineal poblacional es distinto de cero a un nivel de significancia del 5%

4. En el siguiente cuadro se presentan los datos sobre el diámetro ecuatorial (mm) de 30 frutos de naranjo y el peso (grs) de los mismos en el momento de la cosecha.

Fruto	Diámetro (x_1)	Peso (x_2)	Fruto	Diámetro (x_1)	Peso (x_2)
1	69.15	199	16	68.9	171
2	64.43	150	17	72.05	205
3	69.31	190	18	71.46	198
4	69.4	179	19	67.34	171
5	66.3	163	20	67.14	153
6	62.83	143	21	64.91	154
7	63.3	142	22	73.87	212
8	70.03	189	23	66.92	174
9	75.67	236	24	70.3	185
10	61.41	132	25	71.27	210
11	64.87	152	26	68.51	176
12	72.58	176	27	75.47	222
13	66.31	160	28	67.03	170
14	64.24	136	29	73.76	216
15	64.36	148	30	76.17	225

a) Construya un diagrama de dispersión.
b) Calcule el coeficiente de correlación lineal de Pearson en la muestra.
c) Realice una prueba de hipótesis para probar que el coeficiente de correlación lineal poblacional es distinto de cero a un nivel de significancia del 5%
d) Compare sus resultados con los obtenidos en Infostat:



Correlación de Pearson

Variable(1)	Variable(2)	n	Pearson	p-valor
Peso (grs)	Diámetro (mm)	30	0.95	<0.0001

UNIDAD VII
ANALISIS DE REGRESION LINEAL SIMPLE

7.1 INTRODUCCION

Existen situaciones en las cuales el investigador desea verificar la relación funcional que eventualmente puede existir entre dos variables cuantitativas. Así, por ejemplo: cuando **X** es la cantidad de fertilizante y **Y** la producción de caña (TCH = toneladas de caña por hectárea); **X** el peso al nacer de lechones y **Y** el peso a los 30 días de nacidos; **X** el diámetro a la altura del pecho y **Y** la altura de árboles de *Pinus maximinoii*; **X** el año y **Y** la producción de maíz obtenida en cada uno de estos años; **X** la variable tiempo (minutos, por ejemplo) y **Y** la velocidad de infiltración del agua en un tipo determinado de suelo. La variable **X** es conocida como independiente o regresora, y por lo regular considerada como fija y predeterminada, en tanto que la variable **Y** es denominada dependiente, y por lo regular considerada como aleatoria.

A continuación se estudiará la relación de tipo lineal, esto es, los casos en los cuales una variable dependiente **Y** puede ser descrita como una función lineal de una variable independiente **X**. La recta obtenida se denomina: **recta de regresión lineal "y" sobre "x".**

7.2 LEY MATEMÁTICA Y LEY ESTADÍSTICA

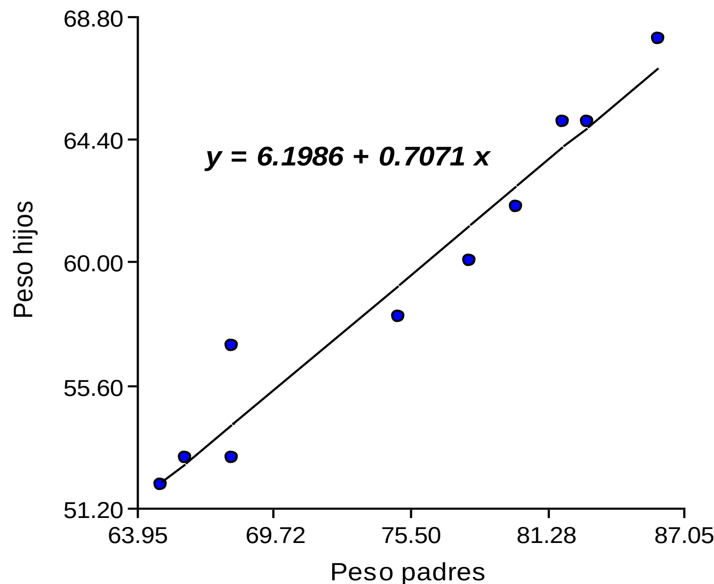
Un hecho que se resaltará desde el inicio es la diferencia conceptual entre una ley matemática y una ley estadística: cuando en un estudio teórico decimos, por ejemplo, que y = 6.1968 + 0.7071 x, estamos diciendo que para cualquier x, el valor correspondiente de y está siempre sobre la recta cuya ecuación es y = 6.1968 + 0.7071 x.

Ejemplo 111

En la tabla siguiente se presentan los pesos de los padres (X) y de sus hijos (Y) en kilogramos.

Par	x_i	y_i	x_i^2	y_i^2	$x_i\,y_i$
1	78	60	6084	3600	4680
2	65	52	4225	2704	3380
3	86	68	7396	4624	5848
4	68	53	4624	2809	3604
5	83	65	6889	4225	5395
6	68	57	4624	3249	3876
7	75	58	5625	3364	4350
8	80	62	6400	3844	4960
9	82	65	6724	4225	5330
10	66	53	4356	2809	3498
	751	593	56947	35453	44921

En este caso podemos decir que el peso Y de los hijos es una función lineal del peso X de sus padres, y estaremos admitiendo el hecho que una nube de puntos descritos en un diagrama de dispersión, puede ser descrita de forma adecuadamente aproximada por una recta cuya ecuación es y = 6.1986 + 0.7071 x, tal como se muestra en la figura siguiente:



Con esto, queremos decir que, cualquiera que sea el valor de X considerado, digamos X = 3, habrá en correspondencia un valor de Y obtenido a través de y = f(x) = 6.1986 + 0.7071 *x*, que llamaremos de ***y* calculado** o ***y* estimado** y que se denota por $\hat{y}$, tal que el par (x, $\hat{y}$), en el ejemplo (3, 8.3199), estará siempre exactamente sobre la recta. De esta forma, el ajuste de y será absolutamente correcto y exento de errores, a través de la función dada. En otras palabras, para ajustar el valor de cualquier y_i de un par ordenado (x_i, y_i) basta que se utilice la función $\hat{y}$ = 6.1986 + 0.7071 x_i y obtendremos el par $(x_i, \hat{y}_i)$ con $y_i = \hat{y}_i$, o sea, si $\hat{e}_i = y_i - \hat{y}_i$ entonces $\hat{e}_i = 0 \;\; \forall_i$.

Por otra parte, en el caso de la ley estadística la estimación no está, por lo general, exenta de errores. Para interpretar bien esta idea, vamos a retomar el diagrama de dispersión, que contiene la recta de la ecuación $\hat{y}$ = 6.1986 + 0.7071 x_i

Así, si el investigador juzga adecuado describir sus datos experimentales a través de la función $\hat{y}$ = 6.1986 + 0.7071 x_i, él deberá estar consciente de que habrá errores de ajuste del valor obtenido (ajustado o estimado) a través de esa función en relación con los valores realmente observados en el experimento.

Vea, por ejemplo, en la tabla 1, que para el primer par los valores observados fueron: peso del padre x_i =78 kg y el peso del hijo y_i =60 kg. Y calculando el valor de $\hat{y}_i$ a través de la función, tenemos: $\hat{y}$ = 6.1986 + 0.7071 (78) = 61.35, que nos lleva a un error de ajuste dado por la diferencia entre el valor observado (colectado, medido) durante el experimento y el correspondiente valor que la función ajusta, dado por: $\hat{e}_i = y_i - \hat{y}_i$ =60.00 – 61.35 = –1.35.

7.3 LA RECTA DE MINIMOS CUADRADOS

Sea $y_i = \alpha + \beta x_i$ la función que queremos ajustar a los datos y el error de ajuste dado por la diferencia entre el valor observado (colectado, medido) durante el experimento y el correspondiente valor que la función ajusta, por: $\hat{e}_i = y_i - \hat{y}_i = y_i - \alpha - \beta x_i$
Entonces, la suma de los cuadrados de los desvíos (e_i) para todos los puntos es dada por:

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right)^2$$

Para minimizar esa suma, cuando varían α y β, debemos igualar a cero las derivadas parciales $\frac{\partial D}{\partial \alpha}$ y $\frac{\partial D}{\partial \beta}$

dadas por: $\frac{\partial D}{\partial \alpha} = 2\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)(-1)$ y $\frac{\partial D}{\partial \beta} = 2\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)(-x_i)$,

que al igualarlas a cero resulta:

(1) $-2\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta} x_i\right) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i = 0$

(2) $-2\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)(-x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i - \hat{\alpha}\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} x_i^{\,2} = 0$

De (1) se obtiene: $\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}\frac{\sum_{i=1}^{n} x_i}{n}$, que es equivalente a : $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$,

y substituyendo $\hat{\alpha}$ en (2) tenemos:

$$\sum_{i=1}^{n} x_i y_i - \left(\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}\frac{\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} x_i^{\,2} = 0$$

$$\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} + \hat{\beta}\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} - \hat{\beta}\sum_{i=1}^{n} x_i^{\,2} = 0$$

$$\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = \hat{\beta}\sum_{i=1}^{n} x_i^{\,2} - \hat{\beta}\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

, que al despejar, da la ecuación siguiente:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Entonces, la recta de mínimos cuadrados o recta de regresión lineal simple es dada por:

$$\hat{y}_i = \hat{\alpha} \pm \hat{\beta} x_i \ .$$

Siendo: $\hat{\alpha}$ y $\hat{\beta}$, respectivamente los estimadores de mínimos cuadrados de los parámetros poblacionales α (coeficiente de posición o intercepto) y β (coeficiente de regresión lineal). Para el ejemplo de los pesos de los padres (X) y de sus hijos (Y), tenemos, con base en la tabla 2.

$$\hat{\beta} = \frac{44921 - \frac{(751)\,(593)}{10}}{56947 - \frac{(751)^2}{10}} = \frac{386.7}{546.9} = 0.7071$$

$$\hat{\alpha} = 59.3 - (0.7071)\,(75.1) = 6.1968$$

Entonces la ecuación de la recta de regresión lineal o de la recta de mínimos cuadrados o de la mejor recta ajustada a esos datos es: $\hat{y} = 6.1986 + 0.7071\, x_i$

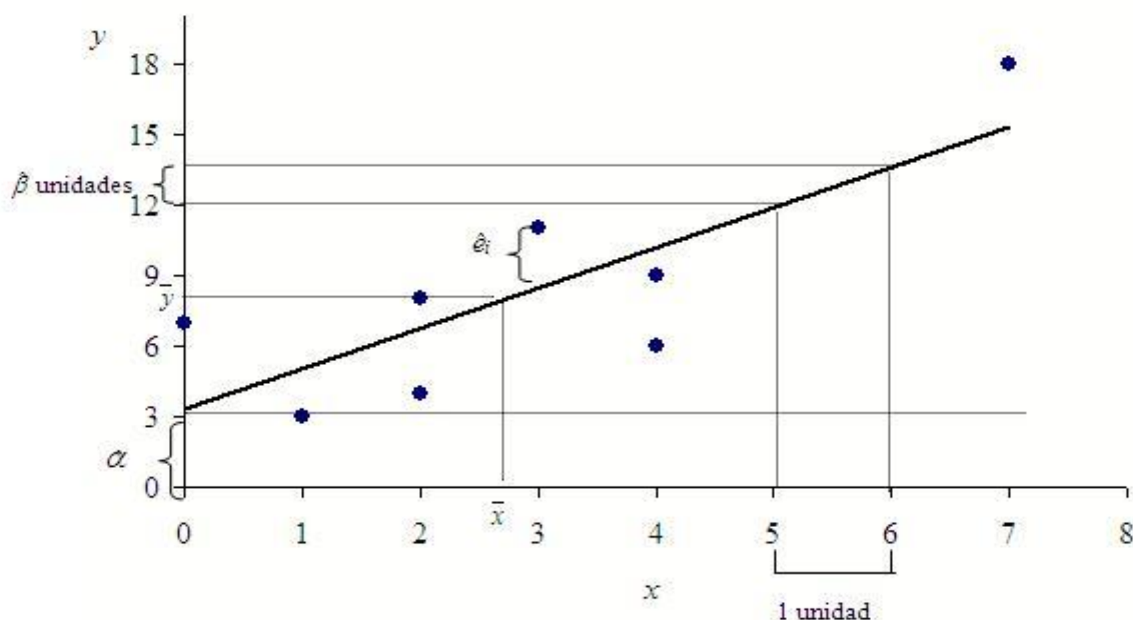
IMPORTANTE:

a) El coeficiente de posición o intercepto (α o β_0), indica la posición en la cual la recta corta el eje **Y.** Si la recta pasa por el origen, entonces α =0. En términos prácticos, indica el valor que asume la variable Y cuando la variable es X=0.

b) El coeficiente de regresión lineal o coeficiente angular de la regresión (β o β_1), determina la pendiente de la recta. Este coeficiente indica la variación en Y causada por la variación de **una unidad en X.** Para el ejemplo que venimos trabajando, por cada incremento de un kilogramo en el peso de los padres, existe un incremento de 0.7071 kg en el peso de los hijos.

c) La recta de regresión pasa por el punto (x, y), esto es, cuando $x = \bar{x}$ tenemos que $\hat{y} = y$.

En la Figura que se presenta a continuación se ilustran las ideas básicas sobre la ecuación ajustada.



7.4 EL MODELO DE REGRESIÓN LINEAL SIMPLE (MRLS) EN FORMA MATRICIAL

Vimos que la muestra aleatoria bajo el modelo de regresión lineal es dada por:

$$y_1 = \beta_0 + \beta_1 + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 + \varepsilon_2$$

$$\cdot \quad \cdot \quad \cdot$$

$$y_n = \beta_0 + \beta_1 + \varepsilon_n$$

$\varepsilon_i \sim N(0 ; \sigma^2)$, $Cov[\varepsilon_i , \varepsilon_j] = 0$, $i \neq j$ $_{i,j} = 1, 2, ... , n$

β_0 y β_1, son constantes desconocidas

$x_1, . . . , x_n$, son constantes conocidas.

Vamos a expresar este modelo usando notación matricial. Sean los vectores:

$$\vec{y} = {}_n\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}_1 , \qquad \vec{\varepsilon} = {}_n\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}_1 \qquad y \qquad \vec{\theta} = {}_2\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_1 ,$$

y sea la matriz X, denominada matriz de modelo o de diseño:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}_{n \times 2},$$

note que el número de columnas de X es igual al número de elementos de θ y el número de líneas es el tamaño de la muestra. La primera columna de X es un vector con los valores que multiplican β_0, por tanto, un vector con elementos iguales a 1. La segunda columna de X es un vector con los valores que multiplican β_1, por tanto, los valores $x_1, . . . , x_n$.

Entonces,

$$X\theta + \varepsilon = Y, \rightarrow \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \rightarrow \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ . \quad . \quad . \quad . \\ . \quad . \quad . \quad . \\ . \quad . \quad . \quad . \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}.$$

Siendo:

Y	=	vector de las observaciones
X	=	matriz del modelo
θ	=	vector de parámetros
ε	=	vector de los residuos o errores

A través de la solución de mínimos cuadrados podemos estimar los parámetros del modelo de regresión lineal simple: $\hat{\theta} = (X'X)^{-1} X'Y$

Ejemplo 112

Considere el caso en que X = dosis de hormona (en UI) y Y = incremento de peso de pollos de engorde (en libras). En forma matricial se presentan los siguientes resultados:

$$Y = \begin{bmatrix} 1.35 \\ 1.42 \\ 1.64 \\ 1.80 \\ 1.94 \\ 2.00 \end{bmatrix}_{6 \times 1} \qquad X = \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 25 \\ 1 & 30 \\ 1 & 35 \end{bmatrix}_{6 \times 2} \qquad \vec{\theta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}_{2 \times 1}$$

$$X'X = {}_{2}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 15 & 20 & 25 & 30 & 35 \end{bmatrix}_{6} \; {}_{6}\begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 25 \\ 1 & 30 \\ 1 & 35 \end{bmatrix}_{2} = {}_{2}\begin{bmatrix} 6 & 135 \\ 135 & 3475 \end{bmatrix}_{2}$$

Para calcular $(X'X)^{-1}$, considere la siguiente matriz cuadrada:

$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, entonces: $A^{-1} = \dfrac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$. Entonces:

$$(X'X)^{-1} = \frac{1}{(6)(3475) - (135)(135)} \; {}_{2}\begin{bmatrix} 3475 & -135 \\ -135 & 6 \end{bmatrix}_{2} = \frac{1}{2625} \; {}_{2}\begin{bmatrix} 3475 & -135 \\ -135 & 6 \end{bmatrix}_{2} = {}_{2}\begin{bmatrix} 1.3238 & -0.0514 \\ -0.0514 & 0.0023 \end{bmatrix}_{2}$$

Para calcular ($X'Y$) se realiza la siguiente operación:

$$X'Y = {}_{2}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 15 & 20 & 25 & 30 & 35 \end{bmatrix}_{6} \; {}_{6}\begin{bmatrix} 1.35 \\ 1.42 \\ 1.64 \\ 1.80 \\ 1.94 \\ 2.00 \end{bmatrix}_{1} = {}_{2}\begin{bmatrix} 10.15 \\ 240.8 \end{bmatrix}_{1} .$$

Finalmente se tiene que: $\hat{\theta} = (X'X)^{-1} X'Y$, es igual a:

$$\hat{\theta} = {}_{2}\begin{bmatrix} 1.3238 & -0.0514 \\ -0.0514 & 0.0023 \end{bmatrix}_{2} \; {}_{2}\begin{bmatrix} 10.15 \\ 240.8 \end{bmatrix}_{1} = {}_{2}\begin{bmatrix} 1.05266 \\ 0.0284 \end{bmatrix}_{1}$$

La ecuación del modelo se escribe: Incremento de peso = $1.05266 + 0.0284 \times$ Dosis

7.5 SUPUESTOS DEL MODELO DE REGRESIÓN

Al efectuar un análisis de regresión se comienza proponiendo una hipótesis acerca del modelo adecuado de la relación entre la variable dependiente y la (s) independiente (s). Para el caso de la regresión lineal simple, el modelo de regresión supuesto es: $y_i = \alpha + \beta x_i + \varepsilon_i$.

Como ya fue explicado anteriormente, se aplica el método de los mínimos cuadrados para determinar los valores de $\hat{\alpha}$ y β, que son los estimados de α y β respectivamente, los parámetros del modelo. La ecuación resultante es: $\hat{y}_i = \hat{\alpha} \pm \hat{\beta} x_i$

Posteriormente, un paso importante a realizar consiste en efectuar un análisis de los supuestos del modelo propuesto, lo cual implica determinar el significado (o importancia estadística) de la supuesta relación entre las variables en estudio. Las pruebas de significancia en el análisis de regresión se basan en los siguientes supuestos acerca del término de error ε:

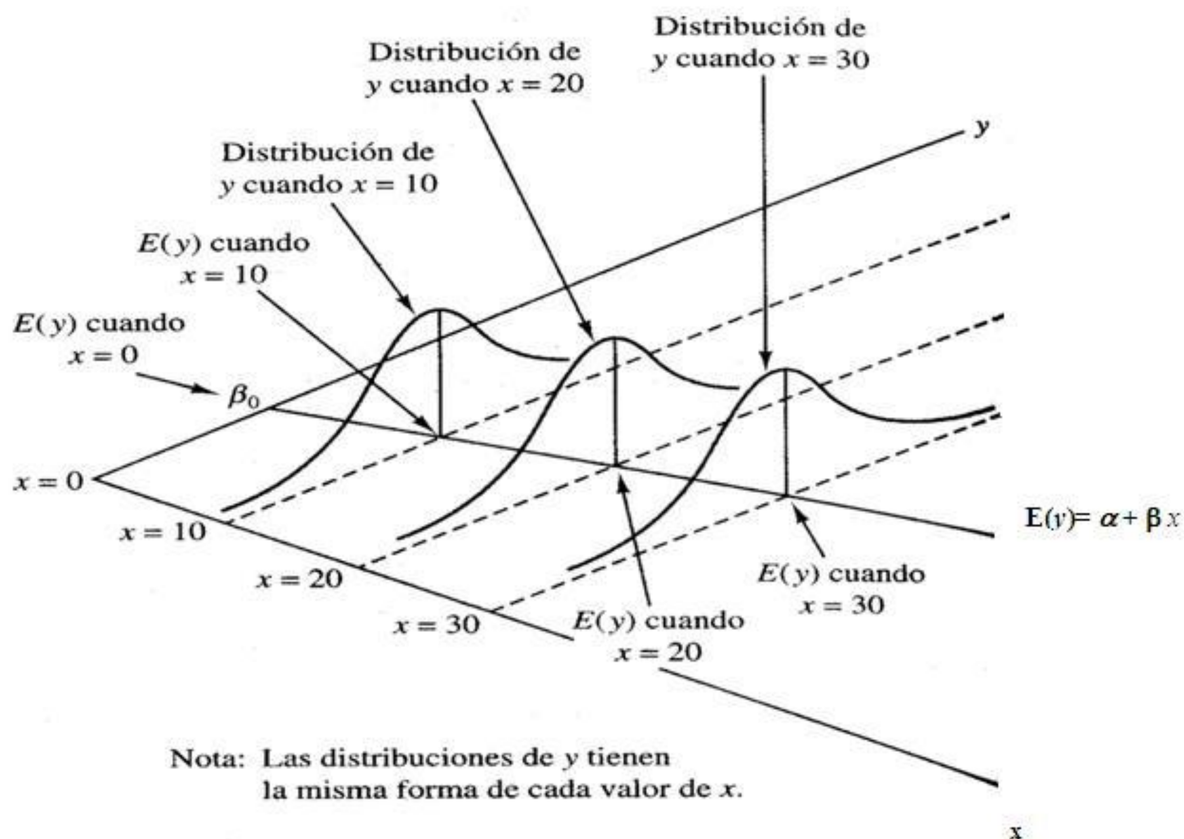
1. El término de error ε es una variable aleatoria con media o valor esperado igual a cero, esto es, E(ε). Esto implica que como α y β son constantes, E(α)= α y E(β)=β. Así, para determinado valor de *x*, el valor esperado de y es:

2. La varianza de ε representada por σ^2, es igual para todos los valores de *x*. Homocedasticidad. Implicación: la varianza de *y* es igual a σ^2, y es la misma para todos los valores de *x*

3. Los valores de ε son independientes.
   Implicación: el valor de ε para un determinado valor de x no se relaciona con el valor de ε para cualquier otro valor de *x*; así, el valor de *y* para determinado valor de *x* no se relaciona con el valor de *y* para cualquier otro valor de *x*.

4. El término de error ε es una variable aleatoria con distribución normal.
   Implicación: como *y* es una función lineal de ε, *y* es también una variable aleatoria distribuida normalmente.

   La siguiente figura ilustra los supuestos del modelo y sus implicaciones:



Observe en la Figura anterior que el valor de E(*y*) cambia de acuerdo con el valor específico de *x* que se considera. Sin embargo, independientemente del valor de *x*, la distribución de probabilidades de ε, y en consecuencia las distribuciones de probabilidades de *y* son normales, y cada distribución tiene la misma varianza. El valor específico del error ε en cualquier punto depende si el valor real de *y* es mayor o menor que E(*y*).

7.6 PRUEBA DE HIPOTESIS SOBRE EL PARÁMETRO β

Con la finalidad de comprobar estadísticamente si las variables **X** y **Y** presentan la supuesta relación lineal, debe realizarse un análisis de varianza (comúnmente abreviado en la literatura como: ANDEVA, ANVA o ANOVA), y evaluar las hipótesis:

Ho : β = 0 (No hay regresión lineal simple)
Ha : β ≠ 0

No rechazar Ho, significa que la pendiente es estadísticamente nula, entonces la recta será paralela al eje **X** y no habrá regresión lineal simple. En otras palabras, en caso de paralelismo, se existe una relación funcional de tipo y = f(x) entre las variables, ella no podrá ser descrita por una ecuación de regresión lineal simple.

A continuación se presenta una tabla con las ecuaciones necesarias para realizar el análisis de varianza. Recuerde que el ANDEVA es un procedimiento aritmético y estadístico que divide la variación total de los datos de **Y** en fuentes de variación, en este caso, una fuente atribuida al modelo de regresión y la otra a la parte no explicada por el modelo (residuo).

Fuentes de variación	**Grados de libertad**	**Suma de Cuadrados (SC)**	**Cuadrados Medios (CM)**	**Valor de la estadística F**
Regresión	p–1	$\hat{\beta} \times \left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \right)$	$\frac{\text{SC Regresión}}{\text{GL Regresión}}$	$\frac{\text{CM Regresión}}{\text{CM Resíduo}}$
Residuo	n-p	SC Total – SC Reg	$\frac{\text{SC Resíduo}}{\text{GL Resíduo}}$	
Total	n-1	$\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i \right)^2}{n}$		

Referencias:

n = cantidad de pares de datos
p = cantidad de parámetros incluidos en el modelo

Para obtener el valor crítico de F, se busca en la tabla de distribución de Fisher-Snedecor el valor de F que está en función de:

n_1 = 1 grado de libertad.
n_2 = n – 2 grados de libertad y,
α = nivel de significancia adoptado.

Regla de decisión: Si Valor de F $\geq F_{crítica}$ rechazar Ho.
O bien analizando el *p–value* (valor p)

Para el ejemplo 111 tenemos:

	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F	Valor crítico de F
Regresión	1	273.43	273.43	149.57	5.32
Residuos	8	14.67	1.83		
Total	9	288.10			

De acuerdo con los resultados del ANDEVA, se concluye que el modelo de regresión lineal simple es adecuado para expresar la relación entre el peso de los padres y el de sus hijos.

7.7 PRUEBA DE HIPÓTESIS ACERCA DE LOS PARÁMETROS DE LA REGRESIÓN LINEAL SIMPLE: USO DE LA PRUEBA t DE STUDENT

7.7.1 Acerca de β

H_0: $\beta = 0$ (No hay relación estadística significativa entre las dos variables)
Ha: $\beta \neq 0$

Estadístico de prueba: $t_o = \dfrac{\hat{\beta} - \beta}{S_{\hat{\beta}}} = \dfrac{0.707 - 0}{0.0578} = 12.23$; solamente en el MRL simple $t_o^2 = F$ ($12.23^2 = 149.57$)

$$S_{\hat{\beta}} = \textit{desviación estimada de } \hat{\beta}$$

$$S_{\hat{\beta}} = \frac{s}{\sqrt{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}} = \frac{1.354}{\sqrt{56,947 - \dfrac{(751)^2}{10}}} = 0.0578$$

$$s = \textit{error estándar estimado}$$

$$s = \sqrt{CMee} = \sqrt{1.834} \quad = 1.354$$

Valor crítico de t:

$$t_{crítico}\left(n-2, \frac{\alpha}{2}\right) = t_{crítico}\left(8, 0.025\right) = 2.306$$

Regla de decisión: Sí $t_o \geq t_{crítico}$ se rechaza la hipótesis nula.

7.7.2 Acerca de α

H_0: $\alpha = 0$ (la recta parte del origen)
Ha: $\alpha \neq 0$

Estadístico de prueba:

$$t_o = \frac{\hat{\alpha} - \alpha}{S_{\hat{\alpha}}} = \frac{6.198 - 0}{4.37} = 1.42$$

$S_{\hat{\alpha}} = desviación\ estimada\ de\ \hat{\alpha}$

$$S_{b_1} = s \times \sqrt{\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}} = 1.354 \times \sqrt{\frac{1}{10} + \frac{75.1^2}{56,947 - \frac{(751)^2}{10}}} = 4.37$$

Valor crítico de t:

$$t_{crítico}\left(n-2, \frac{\alpha}{2}\right) = t_{crítico}\,(8, 0.025) = 2.306$$

Regla de decisión: Sí $t_o \geq t_{crítico}$ se rechaza la hipótesis nula.

En este caso, se concluye que la recta sale del origen.

7.8 COEFICIENTE DE DETERMINACION

El coeficiente de determinación indica la proporción de la variación total que está siendo explicada por la regresión. Además ofrece una idea de la calidad del ajuste del modelo a los datos. El coeficiente de determinación se calcula a través de la siguiente ecuación:

$$R^2 = \frac{SC\ \mathrm{Re}\,gresión}{SC\ Total},\ 0 \leq R^2 \leq 1.$$

Para los datos del ejemplo 111, tenemos que:

$$R^2 = \frac{273.43}{288.10} = 0.949$$

El coeficiente de determinación puede ser expresado en porcentaje, por lo que el valor obtenido anteriormente equivale a 94.9%. Entonces la regresión lineal simple explica 94.9% de la variación total de los datos de peso de los hijos.

OBSERVACIONES:

- El coeficiente de determinación es igual al cuadrado del coeficiente de correlación de Pearson.

- En la regresión puede aplicarse el análisis de correlación para obtener un indicador de la intensidad o fuerza de la relación lineal entre dos variables.

- El valor del coeficiente de determinación debe ser usado con precaución, pues su magnitud depende del número de observaciones en la muestra, tendiendo a crecer cuando n disminuye. Además de eso, es posible volverlo mayor, por la adición de un número suficiente de términos.

- Aunque R^2 aumente sí se adiciona una nueva variable al modelo, esto no significa necesariamente que el nuevo modelo es superior al anterior. A menos que la suma de cuadrados residual del nuevo modelo sea reducida de una cuantía igual al cuadrado medio residual original, el nuevo modelo tendrá un cuadrado medio residual mayor que el original, debido a la pérdida de un grado de libertad. En realidad, ese nuevo modelo podrá ser peor que el anterior.

- La magnitud de R^2 también depende de la amplitud de variación de las variables regresoras (o independientes). Generalmente, R^2 aumentará con mayor amplitud de variación de las X´s y disminuirá en caso contrario. Así, un valor grande de R^2 podrá ser grande simplemente porque los valores de X´s varían en una amplitud muy grande. Por otro lado, R^2 podrá ser pequeño porque las amplitudes de las X´s fueron muy pequeñas para permitir que una relación con Y fuese detectada.

- El R^2 no debe ser considerado en forma aislada para evaluar el ajuste de un modelo de regresión, siempre debe ser acompañado por otros diagnósticos.

- En un intento de corrección de los problema anteriormente señalados, fue definido el coeficiente de determinación ajustado por los grados de libertad, indicado por R^2_{aj}, definido por:

$$R^2_{aj} = R^2 - \frac{1}{n-p}(1-R^2)$$

$$R^2_{aj} = 0.949 - \frac{1}{10-2}(1-0.949)$$

$$R^2_{aj} = 0.9427$$

Excluyendo el caso en que $R^2 = 1$, se tiene que $R^2_{aj} < R^2$.

7.9 INTERVALOS DE (1-α) % DE CONFIANZA

7.9.1 PARA LOS COEFICIENTES

$$\hat{\beta} \pm t_{\left(n-2,\frac{\alpha}{2}\right)} \times S_{\hat{\beta}} = 0.707 \pm 2.306 \times 0.0579$$

$$0.57 \leq \beta \leq 0.84$$

$$\hat{\alpha} \pm t_{\left(n-2,\frac{\alpha}{2}\right)} \times S_{\hat{\alpha}} = 6.198 \pm 2.306 \times 4.37$$

$$-3.88 \leq \alpha \leq 16.28$$

7.9.2 PARA EL VALOR ESTIMADO $\hat{y}_i$

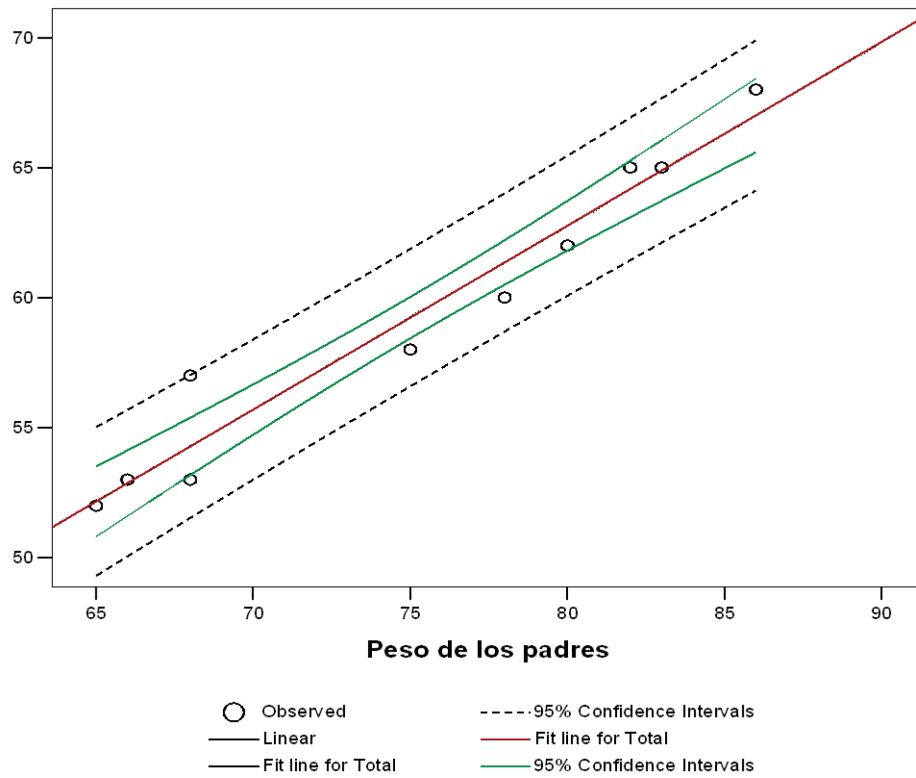
Es un estimado de intervalo del valor medio de *y* para determinado valor de *x*

$$IC\left[\hat{Y}_i\right]_{1-\alpha} = \hat{Y}_i \pm t_{\left(n-2,\frac{\alpha}{2}\right)} \times s \times \sqrt{\frac{1}{n} + \frac{\left(X_i - \bar{X}^2\right)}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}}$$

7.9.3 INTERVALO DE PREDICCIÓN PARA UN VALOR INDIVIDUAL DE y, CUANDO $x = x_p$

$$IC\left[\hat{Y}_p\right]_{1-\alpha} = \hat{Y}_p \pm t_{\left(n-2,\frac{\alpha}{2}\right)} \times s \times \sqrt{1 + \frac{1}{n} + \frac{\left(X_p - \bar{X}^2\right)}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}}$$

En la Figura siguiente, generada en el programa SPSS, se presenta la recta de regresión acompañada de los intervalos de confianza para el valor medio y para la predicción (con líneas punteadas)



Ejemplo 113 **Análisis de regresión lineal simple usando Infostat**

Para los datos del ejemplo 112 se tienen los siguiente resultados:

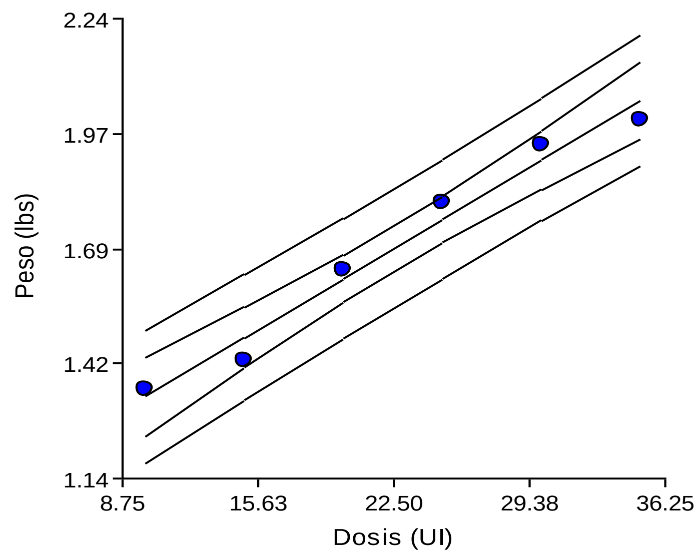


Diagrama de dispersión

Variable	N	R²	R² Aj	ECMP	AIC	BIC
Peso (lbs)	6	0.98	0.97	0.01	-16.11	-16.74

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	1.05	0.05	0.90	1.20	19.49	<0.0001
Dosis (UI)	0.03	2.2E-03	0.02	0.03	12.66	0.0002

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	0.35	1	0.35	160.15	0.0002
Dosis (UI)	0.35	1	0.35	160.15	0.0002
Error	0.01	4	2.2E-03		
Total	0.36	5			

7.10 LIMITACIONES, ERRORES Y ADVERTENCIAS EN EL USO DE LA REGRESIÓN Y EL ANÁLISIS DE CORRELACIÓN

Los análisis de regresión y de correlación son herramientas estadísticas que, cuando se utilizan adecuadamente, pueden ayudar significativamente a tomar decisiones. Pero si se utilizan erróneamente traen como resultado predicciones inexactas y toma de decisiones no deseables. Algunos de los errores más comunes cometidos en el uso de la regresión y correlación se detallan a continuación.

- Extrapolación más allá del intervalo de los datos observados
  Un error común es asumir que la ecuación de estimación puede aplicarse sobre cualquier intervalo de valores. Pero es necesario recordar que una ecuación de regresión es válida solo sobre el mismo intervalo como aquel desde el cual se tomó la muestra inicialmente.

- Causa y efecto
  Otro error que se puede cometer al utilizar el análisis de regresión y correlación es asumir que un cambio en una variable es "ocasionado" por un cambio en la otra variable. Recuerde que: "la regresión y la correlación no pueden determinar la causa y el efecto".

- Uso de tendencias anteriores para estimar tendencias futuras
  Se debe tener cuidado de revaluar los datos anteriores que se utilizan para estimar las ecuaciones de regresión. Las condiciones pueden cambiar y violar una ó más de las suposiciones sobre las cuales depende nuestro análisis de regresión.
  Otro error que puede surgir del uso de datos anteriores se refiere a la dependencia de algunas variables en el tiempo.

- Descubrimiento de relaciones cuando éstas no existen
  Al aplicar el análisis de regresión, las personas algunas veces encuentran una relación entre dos variables que, de hecho no tienen vínculo común. Aun cuando una variable no "ocasiona" un cambio en la otra, piensan que de haber algún factor común a ambas variables. Sería posible por ejemplo, encontrar una relación estadística entre una muestra aleatoria del número de millas por galón consumidas por ocho carros distintos y la distancia de la tierra a cada uno de los otro ocho planetas de nuestro sistema solar. Pero puesto que no existe en absoluto un vínculo común entre el consumo de gasolina y la distancia a otros planetas, esta "relación" no tendría sentido.
  Como en la mayor parte de otras situaciones estadísticas, el investigador debe razonar hacia una conexión entre dos variables antes de trabajar un análisis de regresión.

LISTA DE EJERCICIOS 11

1. Una compañía desea predecir las ventas mensuales a partir de los gastos en publicidad. En primer lugar, se requiere estudiar la relación entre las dos variables: gastos publicitarios (X) y volumen de ventas (Y). En la siguiente tabla se presenta una muestra de los gastos publicitarios y las ventas en los últimos 10 meses.

Mes	Gastos en publicidad En miles de US$	Volumen de ventas en miles de US$
1	1.2	101
2	0.8	92
3	1.0	110
4	1.3	120
5	0.7	90
6	0.8	82
7	1.0	93
8	0.6	75
9	0.9	91
10	1.1	105

Con esta información se le solicita:

a) Construya un diagrama de dispersión.
b) Estime los parámetros de la ecuación de regresión e interpretarlos en términos prácticos.
c) Realice el análisis de varianza y concluya.
d) Calcule el coeficiente de determinación e interpretarlo.
e) Calcule el coeficiente de correlación lineal y pruebe la hipótesis Ho: $\rho = 0$
f) Grafique la recta de regresión en el diagrama de dispersión.
g) ¿Cuál será el volumen de ventas estimado, si se gastan US$1,400.00 en publicidad?

2. En el siguiente cuadro se presentan los datos referentes a la altura (expresada en metros), el diámetro a la altura del pecho (expresada en centímetros) y el volumen (m^3/árbol) de 30 árboles de una especie forestal.

No. Árbol	Altura M	DAP Cm	Volumen m^3/arbol	No. árbol	Altura M	DAP Cm	Volumen m^3/arbol
1	15.9	12.4	0.058935	16	20	26.4	0.464594
2	16	14.5	0.112122	17	20.2	25.6	0.441232
3	16.5	16.1	0.142551	18	20.2	27.0	0.490811
4	16.5	21.8	0.261356	19	20.7	22.3	0.343096
5	17.2	15.2	0.13245	20	20.7	26.7	0.491845
6	17.5	16.8	0.164624	21	20.7	29.0	0.580232
7	18.2	22	0.293597	22	20.7	30.7	0.650253
8	18.2	24.5	0.364115	23	21	18.6	0.242148
9	18.6	19.0	0.223798	24	21.2	17.7	0.221369
10	18.7	24.2	0.365012	25	22.2	25.0	0.462454
11	18.7	28.8	0.516966	26	22.5	24.9	0.464961
12	19	18.2	0.209764	27	22.5	30.0	0.674932
13	19	22.6	0.323449	28	22.7	20.2	0.308719
14	19	27.6	0.4824	29	23	25.3	0.490687
15	19.4	21.0	0.285151	30	23.2	30.0	0.69593

a) Construya un diagrama de dispersión entre las variables DAP (x) y Altura (y).

b) Estime los parámetros de la ecuación de regresión e interprételos en términos prácticos.

c) Realice el análisis de varianza y concluya.

d) Calcule el coeficiente de determinación e interprételo.

e) Grafique la recta de regresión en el diagrama de dispersión.

f) Calcule los residuos y grafíquelos, analice su comportamiento.

3. En una evaluación de un bosque natural, se midió el diámetro a la altura del pecho (DAP) de los árboles dentro de las parcelas, por lo costoso que es la toma de datos de altura, solo se midió la altura comercial (Hc) de algunos árboles, con el fin de obtener una regresión lineal para inferir los valores de altura comercial del total de los árboles dentro de las parcelas. Los datos obtenidos fueron los siguientes:

DAP (cm)	20	45	60	35	42	56	34	28	25	40
Hc (m)	16	22	24	19	20	22	19	17	18	20

Con estos datos, realice el análisis de regresión lineal simple y discuta los resultados.

4. De acuerdo con lo reportado en el Boletín Estadístico del Centro Guatemalteco de Investigación y Capacitación de la Caña de Azúcar (CENGICAÑA), el rendimiento de azúcar, expresado en toneladas métrica por hectárea (TAH) para la zafra 2006/2007 fue de 10.54. Los datos correspondientes a la serie histórica, 1959/60 a 2006/07 se presentan a continuación:

Zafra	**TAH**	**Zafra**	**TAH**	**Zafra**	**TAH**	**Zafra**	**TAH**
1959-60	5.24	1971-72	6.46	1983-84	6.77	1995-96	7.85
1960-61	5.25	1972-73	5.94	1984-85	6.55	1996-97	9.04
1961-62	5.21	1973-74	6.92	1985-86	7.28	1997-98	9.89
1962-63	5.68	1974-75	7.11	1986-87	7.11	1998-99	8.83
1963-64	5.54	1975-76	7.1	1987-88	6.75	1999-00	9.56
1964-65	5.32	1976-77	6.63	1988-89	6.73	2000-01	9.56
1965-66	5.42	1977-78	6.54	1989-90	7.63	2001-02	10.4
1966-67	5.82	1978-79	6.76	1990-91	8.12	2002-03	9.98
1967-68	6.11	1979-80	6.02	1991-92	8.27	2003-04	10.38
1968-69	6.12	1980-81	5.74	1992-93	7.86	2004-05	10.45
1969-70	5.72	1981-82	7	1993-94	7.95	2005-06	10.04
1970-71	6.45	1982-83	7.23	1994-95	8.62	2006-07	10.54

Ajuste y evalúe un modelo de regresión lineal simple. Pronostique el rendimiento de azúcar para la zafra 2007/2008.

5. Ajuste y evalúe el modelo de regresión lineal simple para expresar la relación entre el tiempo y la el rendimiento de caña (expresado en toneladas métrica por hectárea) TCH, para la agroindustria azucarera de Guatemala. Pronostique el rendimiento de caña para la zafra 2007/2008.

Zafra	TCH	Zafra	TCH	Zafra	TCH	Zafra	TCH
1959-60	54	1971-72	71.08	1983-84	72.71	1995-96	78.99
1960-61	57.38	1972-73	72.16	1984-85	66.3	1996-97	88.21
1961-62	55.7	1973-74	78.98	1985-86	70.33	1997-98	97.49
1962-63	60.19	1974-75	81.09	1986-87	72.88	1998-99	87.4
1963-64	59.48	1975-76	82.29	1987-88	73.33	1999-00	82.8
1964-65	56.83	1976-77	78.93	1988-89	70.06	2000-01	84.64
1965-66	62.06	1977-78	78.94	1989-90	80.32	2001-02	92
1966-67	63.65	1978-79	78.99	1990-91	82.79	2002-03	88.32
1967-68	63.43	1979-80	70.07	1991-92	80.02	2003-04	91.89
1968-69	64.56	1980-81	70.33	1992-93	77.92	2004-05	91.3
1969-70	61.9	1981-82	83.29	1993-94	77.49	2005-06	89.3
1970-71	67.75	1982-83	75.26	1994-95	86.11	2006-07	96.31

Fuente: Boletín Estadístico, Año 8, No.1, Noviembre de 2007. CENGICAÑA. Disponible en: www.cengicana.org

6. Examine los datos siguientes, referentes a la variación del porcentaje de ventas de pintura para casas (PV) cuando se aplican diferentes porcentajes de descuento (PD): (Valor 50 puntos)

PD	10	12	14	16	18	20	24	26	28	30
PV	3.43	3.75	4.52	5.13	5.94	6.35	7.99	8.23	9.46	10.35

El contador jefe de la empresa confía en un informe nítido y organizado. Por consiguiente, deberán analizarse los datos en el orden siguiente:

a) De una idea de la relación entre las dos variables utilizando un diagrama de dispersión.
b) Obtenga la ecuación de mínimos cuadrados e interprete los parámetros.
c) Mediante el empleo de análisis de varianza, determine qué porcentaje de la variación es explicada por la ecuación de regresión estimada.
d) ¿Puede rechazar la hipótesis que el porcentaje de descuento no afecta la variación en las ventas (Use α=0.05)? Explique.
e) Estime el porcentaje de variación en las ventas, cuando el porcentaje de descuento es de 35.
f) Cuando el valor de porcentaje de ventas es de 7, cuál es el valor de descuento estimado?

7. Una investigación de la relación entre el flujo de tránsito (x) (miles de automóviles por día) y el contenido de plomo (Y) de la corteza de árboles cerca de la autopista (μg/g) de peso en seco produjo los siguientes datos:

Día	1	2	3	4	5	6	7	8	9	10	11
x	8.3	8.3	12.1	12.1	17.0	17.0	17.0	24.3	24.3	24.3	33.6
Y	227	312	362	521	640	539	728	945	738	759	1263

a) Construya un diagrama de dispersión.
b) Ajuste un modelo de regresión lineal, interprete los parámetros estimados en términos prácticos.

c) Evalúe la relación entre flujo de tránsito y el contenido de plomo por medio del análisis de varianza. Use un nivel de significancia del 5%.
d) Calcule el coeficiente de determinación e interprételo.
e) Cuando el flujo de tránsito es de 30,000 automóviles por días, cuál es el valor esperado del contenido de plomo en los árboles?
f) Si el contenido esperado de plomo de la corteza de árboles cerca de la autopista es de 400 (μg/g), cuál es el flujo estimado de automóviles?

8. Los diámetros y volúmenes de los árboles de una parcela de 576 m^2 situada en una plantación de paraíso gigante (*Melia azedarach* var. gigantea) de 8 años en Santiago del Estero, Argentina, se presentan a continuación:

Dap(cm)	Vol(m^3)	Dap (cm)	Vol(m^3)	Dap (cm)	Vol(m^3)
9.50	0.03	15.20	0.07	17.20	0.09
11.90	0.04	15.50	0.07	17.20	0.09
12.00	0.04	16.00	0.08	17.30	0.09
12.80	0.05	16.20	0.08	17.50	0.10
13.20	0.05	16.40	0.08	18.30	0.10
13.30	0.05	16.60	0.09	19.10	0.11
13.60	0.06	16.70	0.09	19.20	0.12
14.20	0.06	16.90	0.09	19.50	0.12
14.40	0.06	17.00	0.09	21.20	0.14
15.00	0.07	17.00	0.09	21.50	0.15

Con estos datos realice un análisis de regresión lineal simple para estudiar la posible relación entre el DAP y el volumen de los árboles.

7.11 OTROS MODELOS DE REGRESIÓN

Por ahora, solamente se han considerado modelos con una variable predictora. La idea es tratar de aumentar la medida de ajuste R^2 del modelo, sin incluir variables predictoras adicionales. Lo primero que hay que hacer es un diagrama de dispersión para observar el tipo de tendencia. Pueden resultar gráficos como los que aparecen en las Figuras 1a y 1b.

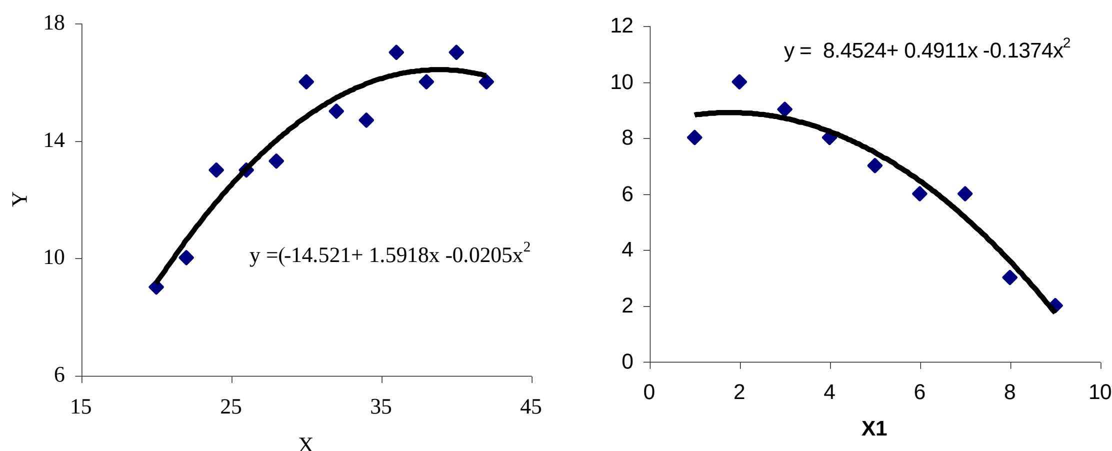


Figura 1. Gráficos de un modelo cuadrático.

El modelo cuadrático o polinomial de grado 2, que es de la forma general $y = a + bx + cx^2$ puede ser modelado como una regresión múltiple con dos variables predictoras. La Figura 2a y 2b corresponde a un modelo **EXPONENCIAL** de la forma $y = \alpha\, e^{\beta x}$ con α positivo y β negativo y positivo (respectivamente). Este modelo es muy adecuado para modelar crecimientos poblacionales.

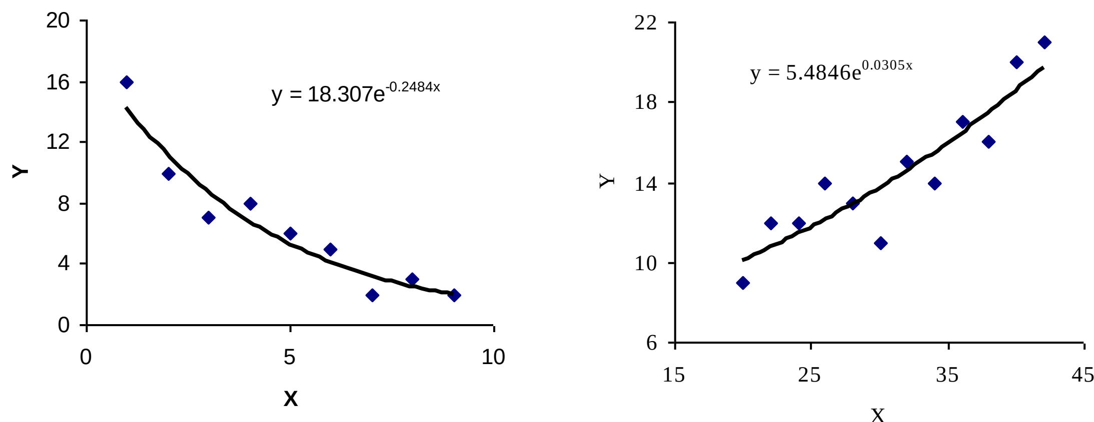


Figura 2. Gráficos de un modelo exponencial

La tercera figura corresponde a un modelo POTENCIA (o doblemente logarítmico) de la forma $y = \alpha\, x^{\beta}$

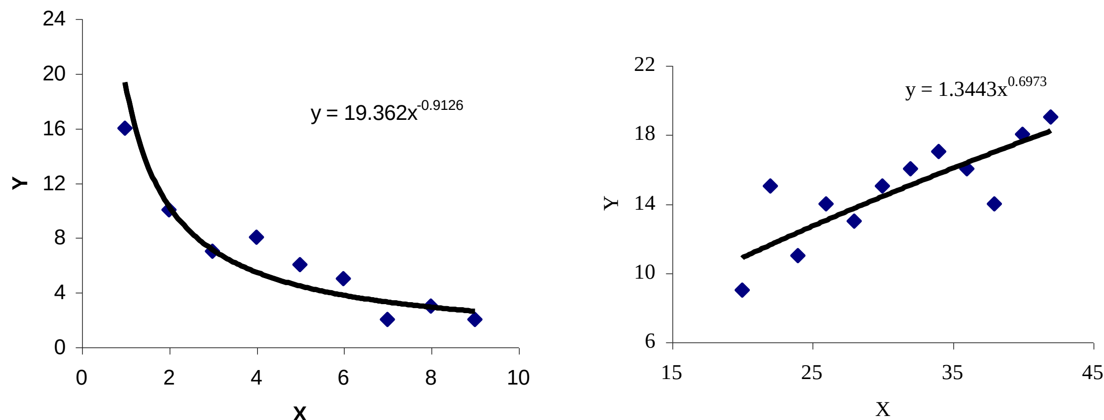


Figura 3a Modelo potencia negativo　　　Figura 3b Modelo potencia positivo

A continuación se presentan gráficas de otros modelos de regresión.

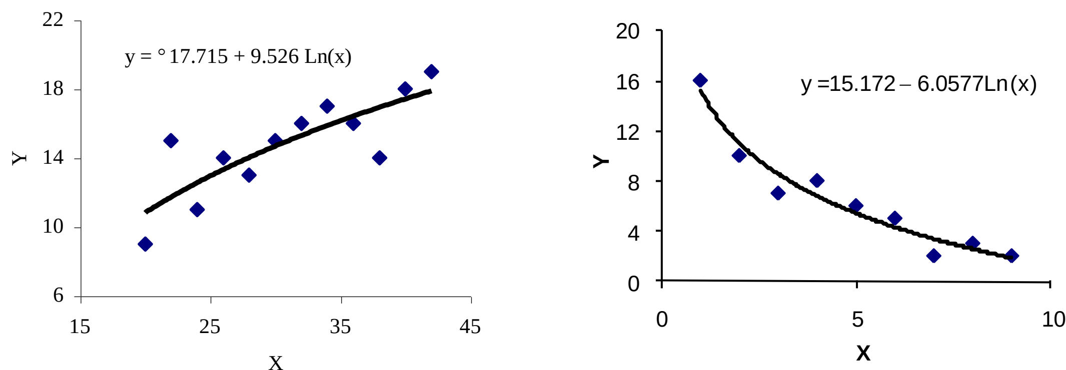


Figura 4. Modelo logarítmico

La tabla siguiente muestra las transformaciones de la variable predictora y/o respuesta que se requieren para linealizar varios modelos.

Nombre del modelo	Ecuación	Transformación	Modelo linealizado
Exponencial	$y = \alpha\, e^{\beta x}$	Z= ln(y) x=x	$Z = \ln \alpha + \beta x$
Potencia o doblemente logarítmico (*)	$y = \alpha\, x^{\beta}$	Z= ln(y) W=ln(x)	$Z = \ln \alpha + \beta W$
Logarítmico (**)	$y = \alpha + \beta \ln (x)$	Y=y W = ln (x)	$y = \alpha + \beta W$
Geométrico	$y = \alpha\, \beta^{x}$	Z=ln (y)	$Z = \ln \alpha + x \ln (\beta)$
Inversa o hiperbólica	$y = \alpha + \beta\, 1/x$	Y=Y W =1/x	$y = \alpha + \beta W$
Doblemente inversa	$y =1/(\alpha + \beta x)$	Z = 1/y x = x	$Z = \alpha + \beta x$

(*) Algunos autores se refieren a este modelo como logarítmico.

(**) También referido como semilogarítmico

Nota:
El primero, el segundo modelo y el cuarto modelo son válidos bajo la suposición de que los errores son multiplicativos y habría que evaluar esta suposición, haciendo análisis de residuos, si los logaritmos de los errores tienen una media de cero y una varianza constante. Si los errores no son multiplicativos entonces deberían aplicarse técnicas de regresión no lineal.

LISTA DE EJERCICIOS 12

1. A continuación se presentan las mediciones de DAP (cm) y ALTURA (m) de 10 árboles tipo de *Pinus oocarpa*:

No.	DAP (X)	ALTURA (Y)
1	34.0	18.5
2	25.5	21.0
3	16.5	15.3
4	17.0	17.8
5	25.0	16.8
6	12.0	14.5
7	23.5	14.2
8	13.0	9.2
9	25.8	16.0
10	17.6	15.5

Con estos datos evalúe las ecuaciones siguientes para estimar la altura:

a) $\text{Altura} = \hat{\beta}_0 + \hat{\beta}_1 \ln \text{DAP}$ (logarítmica)

b) $\text{Altura} = 1.3 + \hat{\beta}_0 \text{ DAP} + \hat{\beta}_1 \text{ DAP}^2$; $\hat{Y} = (\text{Altura} - 1.3) / \text{DAP}$ (Ecuación parabólica)

c) $\text{Altura} = 1.3 + \left(\dfrac{\text{DAP}}{\hat{\beta}_0 + \hat{\beta}_1 \text{ DAP}} \right)^2$; $\hat{Y} = \text{DAP} \sqrt{\text{ALTURA} - 1.3}$ (Ecuación de Nasslund)

2. Pedro Agustín López Velásquez (2003) realizó el trabajo de tesis titulado "Estudio del crecimiento y rendimiento de *Pinus pseudostrobus* Lindley, en bosques naturales de los departamentos de Chimaltenango y Sololá". Evaluó las relaciones: edad–altura, edad–dap y edad–volumen, en 3 calidades de sitio; utilizando los modelos de regresión siguientes:

1. $Y = b_o + b_1 \dfrac{1}{X^k}$ (Modelo de Schumacher), siendo k = 0.597844

2. $Y = e \;\; ((b_o + b_1 \ln X + b_2 (\ln X)^2)$

3. $Y = \dfrac{X^2}{b_o + b_1 X + b_2 X^2}$

4. $Y = b_o + b_1 \ln X$

5. $Y = b_o + b_1 X + b_2 X^2$

6. $Y = \dfrac{X^2}{b_o + b_1 X^2}$

7. $\ln Y = b_o + b_1 \dfrac{1}{X}$

8. $Y = b_o X^{b_1}$

9. $Y = b_o b_1^{X}$

A continuación se presenta tres conjuntos de datos, extraídos de este trabajo de tesis. Para cada relación debe de evaluar los 9 modelos mencionados anteriormente. Seleccione el o los modelos que presenten mejor ajuste.

Árbol	**Edad (años)**	**Altura (m)**
1	1	0.30
2	3	3.07
3	5	5.84
4	7	8.61
5	9	11.20
6	12	15.10
7	15	16.92
8	18	19.69
9	20	21.71
10	23	25.74
11	27	27.76
12	31	29.79
13	33	30.06
14	35	28.65
15	36	29.50
16	41	37.15
17	44	38.18
18	48	40.36
19	52	39.23
20	57	38.20

Árbol	**Edad (años)**	**DAP (cm)**
1	3	3.4
2	6	4.9
3	8	14.3
4	11	11.3
5	13	26.8
6	16	18.5
7	18	40
8	21	27.2
9	23	48.6
10	23	39
11	24	34.5
12	26	33
13	28	54.3
14	31	37.2
15	33	59.5
16	36	41.4
17	44	45.7
18	47	41
19	52	43.4
20	54	53.1

Árbol	**Edad (años)**	**Volumen (m3)**
1	3	0.0027
2	6	0.0111
3	8	0.0872
4	11	0.0654
5	13	0.369
6	16	0.2132
7	18	0.8585
8	21	0.4958
9	23	0.369
10	26	0.9521
11	28	0.8585
12	31	1.3221
13	33	1.4891
14	34	1.6052
15	39	2.1537
16	44	3.1015
17	49	3.7016
18	52	2.7848
19	54	4.0959
20	57	3.1566

3. La velocidad de infiltración es definida como la entrada vertical del agua a través de los poros por unidad de tiempo. La velocidad de infiltración es una de las características del suelo más importantes para el diseño, operación y evaluación de sistemas de riego por aspersión y superficiales, es por esto que se hace necesario obtener información confiable de esta propiedad.

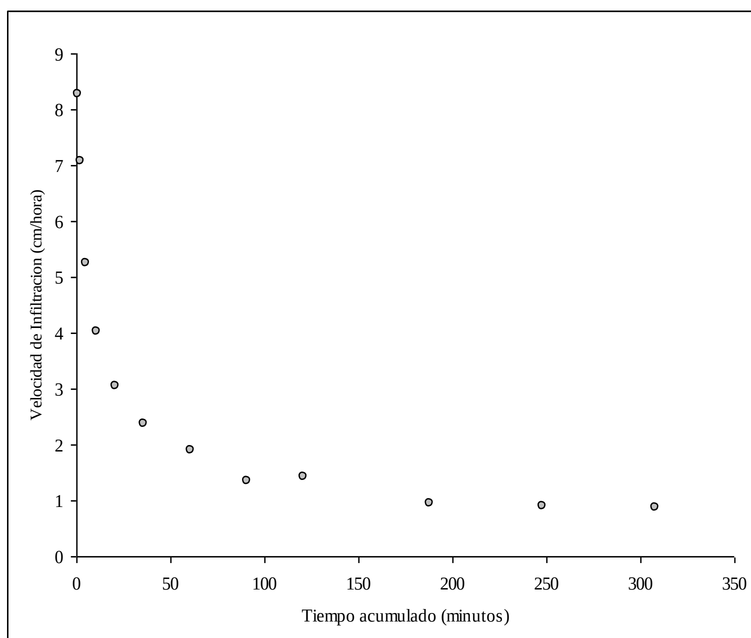
Para el cálculo de la velocidad de infiltración e infiltración acumulada, se han elaborado varios modelos, el más utilizado es el propuesto por Kostiakov–Lewis. Este modelo se basa en que la velocidad de infiltración decrece con el tiempo, siendo representada por la siguiente expresión:

$$I = K\ t^{-n} \quad (y = b_0\ x^{-b_1})$$

Siendo:
I = velocidad de infiltración (cm/hora)
t = tiempo acumulado de infiltración (minutos)
K = parámetro que representa la velocidad de infiltración cuando el tiempo es 1 minuto.
n = parámetro que indica la forma en que la velocidad de infiltración se reduce con el tiempo. Tiene valores entre –1.0 y 0, siendo su valor más común: –0.5.

A continuación se presentan los datos de campo, obtenidos al realizar una prueba de infiltración, utilizando el método del infiltrómetro de doble cilindro. Así como el diagrama de dispersión.

Tiempo acumulado (min)	Velocidad de infiltración (cm/hora)
1	8.28
2	7.08
5	5.26
11	4.02
21	3.054
36	2.372
61	1.899
91	1.354
121	1.416
188	0.961
248	0.9
308	0.88



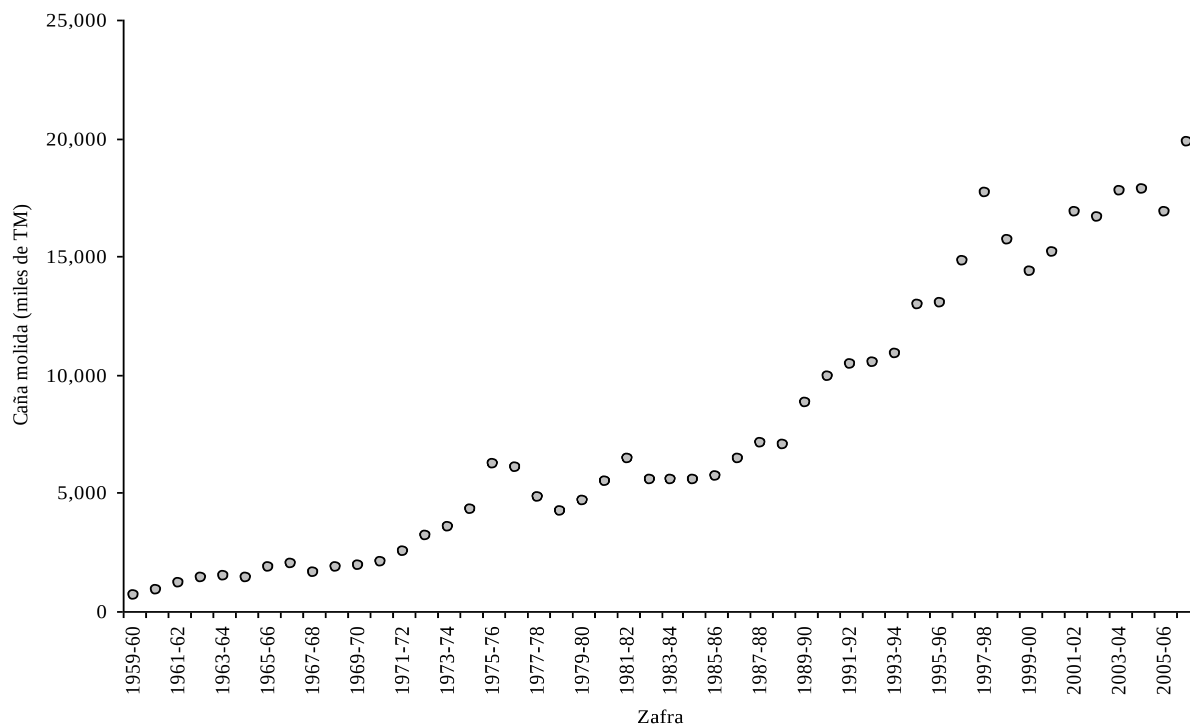
Con estos datos ajuste y evalúe el modelo de Kostiakov–Lewis.

4. De acuerdo con lo reportado en el Boletín Estadístico del Centro Guatemalteco de Investigación y Capacitación de la Caña de Azúcar (CENGICAÑA), la producción nacional de caña molida para la zafra 2006/2007 fue de 19₁813,455 toneladas métricas. Los datos correspondientes a la serie histórica, 1959/60 a 2006/07 se presentan a continuación:

Zafra	Caña molida		Zafra	Caña molida
1959-60	670.13		1983-84	5536.27
1960-61	878.74		1984-85	5569.53
1961-62	1217.47		1985-86	5696.39
1962-63	1373.99		1986-87	6413.25
1963-64	1461.83		1987-88	7113.20
1964-65	1427.07		1988-89	7006.06
1965-66	1844.22		1989-90	8834.89
1966-67	2005.25		1990-91	9934.92
1967-68	1605.11		1991-92	10402.98
1968-69	1852.90		1992-93	10519.42
1969-70	1946.47		1993-94	10847.97
1970-71	2075.29		1994-95	12916.57
1971-72	2543.07		1995-96	13033.51
1972-73	3166.24		1996-97	14792.74
1973-74	3584.44		1997-98	17666.17
1974-75	4258.34		1998-99	15644.72
1975-76	6220.76		1999-00	14338.96
1976-77	6049.35		2000-01	15174.03
1977-78	4785.96		2001-02	16900.24
1978-79	4242.06		2002-03	16623.87
1979-80	4624.55		2003-04	17780.56
1980-81	5485.81		2004-05	17819.76
1981-82	6410.56		2005-06	16883.88
1982-83	5527.19		2006-07	19813.46

Fuente: Boletín Estadístico, Año 8, No.1, Noviembre de 2007. CENGICAÑA. Disponible en: www.cengicana.org

Con estos datos, ajuste un modelo de regresión para expresar la relación entre el tiempo (zafras) y la cantidad de caña molida. Analice la tendencia que se presenta en el gráfico de dispersión:



5. A continuación se presentan los datos de elevación (metros sobre el nivel del mar) y precipitación pluvial anual, reportados en el año 2004 en 36 estaciones ubicadas en la zona cañera de la Costa Sur de Guatemala. Realice un análisis de regresión y explique el comportamiento de la precipitación en función de la altitud sobre el nivel de mar. ¿Cuál es modelo que mejor se ajusta? Investigue sobre la explicación práctica del comportamiento observado.

Estación	Elevación (msnm)	ppt anual
San Luis	5	1496
San Antonio	10	909
Guadalupe	13	1014
Montañesa	21	1331
Amazonas	30	1155
Verapaz	35	1421
La Agrícola	40	1556
Playa Grande	50	1459
La Habana	60	1747
San Patricio	65	1668
Mojarras	69	1805
Agua Blanca	78	2299
Puyumate	85	1933
Naranjales	90	1725
Buenos Aires	96	2195
La Cabaña	110	2688
Santa Marta	115	3360
Belén	142	3402
San Juan Bosco	160	2082
Santa Ana	160	2401
El Refugio	200	3554
Variedades	225	3709
Tululá	229	2786
Torolita	240	3656
San Bonifacio	245	3270
El Bálsamo	275	3891
Camantulul	280	4106
Cengicaña	300	4508
Mangalito	400	4303
Panorama	500	4185
El Baúl	550	3997
Sabana Grande	730	2721
Los Tarros	740	3908
La Eminencia	820	2924
Santa Teresa	1200	1284
El Rincón	1200	983.9

Fuente: Memoria de presentación de resultados de investigación, zafra 2004-2005. CENGICAÑA. Páginas: 152-154.

6. A continuación se presentan los contenidos de nitrógeno en suelo (%N) obtenidos en 3 tipos de áreas con cubierta forestal en una microrregión del Estado de Amazonas (Brasil): PP = con Paricá (*Schizolobium excelsum* var. amazonicum Ducke); PM = Mogno o caoba (*Swietenia macrophylla*) y PN = floresta nativa. En cada área se tomaron muestras en 3 puntos y 9 profundidades (5, 15, 25, 35, 50, 70, 90, 115 y 140 cms).

Área PP		Área PM		Área RN	
Profundidad	%N	Profundidad	%N	Profundidad	%N
5	0.2733135	5	0.2373847	5	0.3141158
15	0.0955491	15	0.1443607	15	0.1457906
25	0.1041372	25	0.0979447	25	0.0937343
35	0.0840927	35	0.0776091	35	0.0826143
50	0.0717175	50	0.0672835	50	0.080701
70	0.0463921	70	0.0477203	70	0.03685245
90	0.0369276	90	0.0423181	90	0.0463487
115	0.0365649	115	0.031948	115	0.0367886
140	0.0295952	140	0.0342743	140	0.0353021
5	0.2439834	5	0.2342962	5	0.2935362
15	0.1126617	15	------------	15	0.1336166
25	0.0930648	25	0.1258502	25	0.1011309
35	0.0766141	35	0.0814077	35	0.0819628
50	0.0620216	50	0.0651261	50	0.061586
70	0.0401482	70	0.0474623	70	0.0412341
90	0.0387405	90	0.0343905	90	0.04055175
115	0.0354053	115	0.028303	115	0.0415754
140	0.036976	140	0.0250555	140	0.0387028
5	0.2373847	5	0.2279078	5	0.341901
15	0.1443607	15	0.1306004	15	0.1726898
25	0.0979447	25	0.0870791	25	0.1260365
35	0.0776091	35	0.0780943	35	0.0933757
50	0.0672835	50	0.0577855	50	0.0793089
70	0.0477203	70	0.0490076	70	0.0627436
90	0.0423181	90	0.0351572	90	0.0476364
115	0.031948	115	0.0315359	115	0.03480715
140	0.0342743	140	0.0269693	140	0.02811045

Ajuste un modelo de regresión para estudiar la relación entre profundidad de muestreo y %N, en cada área. Posteriormente en un gráfico presente las curvas de regresión para las 3 áreas de estudio. ¿Cuáles son sus principales conclusiones?

UNIDAD VIII
ANALISIS DE REGRESION LINEAL MÚLTIPLE

8.1 INTRODUCCIÓN

El análisis de regresión múltiple es el estudio de la forma en que una variable dependiente **Y** se relaciona con dos o más variables independientes (**X**´s). En el caso general se emplea *k* para representar la cantidad de variables independientes.

La ecuación que describe la forma en que una variable dependiente (y por lo regular aleatoria) ***y*** se relaciona con las variables predictoras independiente (fija y predeterminadas, medidas sin error) $X_1, X_2, \ldots, X_k$ y un término de error, se denomina: modelo de regresión, y tiene la forma siguiente:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_k + \varepsilon \quad (1)$$

Siendo que:

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ son los parámetros y ε es el término de error que explica la variabilidad en *y* que no puede explicar el efecto lineal de las *k* variables independientes.

Si se conocieran los valores de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ se podría usar la ecuación (1) para calcular el valor medio de *y* dados los valores de $X_1, X_2, \ldots, X_k$. Desafortunadamente esos parámetros, por lo general, no se conocen y se deben determinar a partir de datos de una muestra. Para calcular los estadísticos de la muestra $b_0, b_1, \ldots, b_p$ que se usan como estimadores puntuales de los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ se usa una muestra aleatoria. Esos estadísticos dan como resultado la siguiente ecuación estimada de regresión múltiple:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p \quad (2)$$

Siendo que: $b_0, b_1, \ldots, b_p$ son los estimadores de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ y $\hat{y}$ es el valor estimado de la variable dependiente.

8.2 ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO USANDO EL MÉTODO MATRICIAL

Para mostrar el procedimiento utilizado para estimar los parámetros de un modelo de regresión lineal múltiple, se utilizará el ejemplo siguiente:

Ejemplo 114

Con la información de la producción de leche promedio por vaca (Y), en cada uno de 20 establos, se desea ajustar un modelo de regresión lineal múltiple, considerando como variables regresoras: X_1 =edad al parto (meses), X_2 = período seco (días) y X_3 = período de lactancia (días). Los datos se presentan en el cuadro siguiente:

Establo	X_1	X_2	X_3	Y
1	26	55	200	4500
2	52	50	275	6280
3	24	40	210	3840
4	61	30	315	7900
5	86	35	360	9100
6	30	65	240	5050
7	28	70	230	4710
8	55	40	280	6300
9	76	30	340	8190
10	60	60	285	6550
11	63	45	305	7870
12	28	65	222	4620
13	27	90	225	3990
14	52	35	290	6800
15	78	30	350	8430
16	49	60	245	5650
17	32	80	235	4900
18	58	40	300	7800
19	54	55	290	6840
20	28	60	250	5750

Recuerde que a través de la solución de mínimos cuadrados podemos estimar los parámetros del modelo de regresión lineal: $\hat{\theta} = (X'X)^{-1} X'Y$. Por lo que tenemos que organizar los datos así:

$$
Y = {}_{20}\begin{pmatrix} 4500 \\ 6280 \\ 3840 \\ 7900 \\ 9100 \\ 5050 \\ 4710 \\ 6300 \\ 8190 \\ 6550 \\ 7870 \\ 4620 \\ 3990 \\ 6800 \\ 8430 \\ 5650 \\ 4900 \\ 7800 \\ 6840 \\ 5750 \end{pmatrix}_{1}
\qquad
X = {}_{20}\begin{vmatrix} 1 & 26 & 55 & 200 \\ 1 & 52 & 50 & 275 \\ 1 & 24 & 40 & 210 \\ 1 & 61 & 30 & 315 \\ 1 & 86 & 35 & 360 \\ 1 & 30 & 65 & 240 \\ 1 & 28 & 70 & 230 \\ 1 & 55 & 40 & 280 \\ 1 & 76 & 30 & 340 \\ 1 & 60 & 60 & 285 \\ 1 & 63 & 45 & 305 \\ 1 & 28 & 65 & 222 \\ 1 & 27 & 90 & 225 \\ 1 & 52 & 35 & 290 \\ 1 & 78 & 30 & 350 \\ 1 & 49 & 60 & 245 \\ 1 & 32 & 80 & 235 \\ 1 & 58 & 40 & 300 \\ 1 & 54 & 55 & 290 \\ 1 & 28 & 60 & 250 \end{vmatrix}_{4}
$$

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 26 & 52 & 24 & 61 & 86 & 30 & 28 & 55 & 76 & 60 & 63 & 28 & 27 & 52 & 78 & 49 & 32 & 58 & 54 & 28 \\ 55 & 50 & 40 & 30 & 35 & 65 & 70 & 40 & 30 & 60 & 45 & 65 & 90 & 35 & 30 & 60 & 80 & 40 & 55 & 60 \\ 200 & 275 & 210 & 315 & 360 & 240 & 230 & 280 & 340 & 285 & 305 & 222 & 225 & 290 & 350 & 245 & 235 & 300 & 290 & 250 \end{pmatrix}$$

4 20

Luego obtenemos:

$$X'X = \begin{pmatrix} 20 & 967 & 1035 & 5447 \\ 967 & 53917 & 45535 & 280166 \\ 1035 & 45535 & 59175 & 270905 \\ 5447 & 280166 & 270905 & 2E+06 \end{pmatrix} \qquad (X'X)^{-1} = \begin{pmatrix} 13.665 & 0.1179 & -0.039 & -0.1 \\ 0.1179 & 0.0022 & 1E-04 & -0 \\ -0.039 & 0.0001 & 4E-04 & 0 \\ -0.064 & -8E-04 & 5E-05 & 0 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 125070 \\ 6605990 \\ 6087900 \\ 3.5E+07 \end{pmatrix} \qquad \theta = \begin{pmatrix} -1186 \\ 9.3929 \\ -7.868 \\ 27.144 \end{pmatrix}$$

8.3 ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO POR MEDIO DE LA SOLUCIÓN DE UN SISTEMA DE ECUACIONES SIMULTÁNEAS

Sea $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$ la función que queremos ajustar a los datos y el error de ajuste dado por la diferencia entre el valor observado (colectado, medido) durante el experimento y el correspondiente valor que la función ajusta, por:

$$\hat{e}_i = Y_i - \hat{y}_i = Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}$$

Entonces, la suma de los cuadrados de los desvíos (e_i) para todos los puntos es dada por:

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} \right)^2$$

Para minimizar esa suma, cuando varían β_0, β_1, y β_2 debemos igualar a cero las derivadas parciales $\frac{\partial D}{\partial \beta_0}$, $\frac{\partial D}{\partial \beta_1}$ y $\frac{\partial D}{\partial \beta_2}$, dadas por:

$$\frac{\partial D}{\partial \beta_0} = 2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}\right)(-1),$$
$$\frac{\partial D}{\partial \beta_1} = 2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}\right)(-X_{1i}),$$
$$\frac{\partial D}{\partial \beta_1} = 2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}\right)(-X_{2i}) .$$

Que al igualarlas a cero resulta:

(1) $-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 X_{1i}-\beta_2 X_{2i}\right)=0 \quad \Rightarrow \quad \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} X_{1i} - \hat{\beta}_2\sum_{i=1}^{n} X_{2i} = 0$

Ecuación 1: $\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n} X_{1i} + \hat{\beta}_2\sum_{i=1}^{n} X_{2i}$

(2) $-2\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1 X_{1i}-\beta_2 X_{2i}\right)\left(X_{1i}\right)=0 \Rightarrow \sum_{i=1}^{n}\left(X_{1i}Y_i-\hat{\beta}_0 X_{1i}-\hat{\beta}_1 X_{1i}{}^2-\hat{\beta}_2 X_{1i}X_{2i}\right)=0$

$\sum_{i=1}^{n} X_{1i}Y_i - \hat{\beta}_0\sum_{i=1}^{n} X_{1i} - \hat{\beta}_1\sum_{i=1}^{n} X_{1i}{}^2 - \hat{\beta}_2\sum_{i=1}^{n} X_{1i}X_{2i} = 0$, finalmente:

Ecuación 2: $\sum_{i=1}^{n} X_{1i}Y_i = \hat{\beta}_0\sum_{i=1}^{n} X_{1i} + \hat{\beta}_1\sum_{i=1}^{n} X_{1i}{}^2 + \hat{\beta}_2\sum_{i=1}^{n} X_{1i}X_{2i}$

(3) $-2\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1 X_{1i}-\beta_2 X_{2i}\right)\left(X_{2i}\right)=0 \Rightarrow \sum_{i=1}^{n}\left(X_{2i}Y_i-\hat{\beta}_0 X_{2i}-\hat{\beta}_1 X_{1i}X_{2i}-\hat{\beta}_2 X_{2i}{}^2\right)=0$

$\sum_{i=1}^{n} X_{2i}Y_i - \hat{\beta}_0\sum_{i=1}^{n} X_{2i} - \hat{\beta}_1\sum_{i=1}^{n} X_{1i}X_{2i} - \hat{\beta}_2\sum_{i=1}^{n} X_{2i}{}^2 = 0$, finalmente:

Ecuación 3: $\sum_{i=1}^{n} X_{2i}Y_i = \hat{\beta}_0\sum_{i=1}^{n} X_{2i} + \hat{\beta}_1\sum_{i=1}^{n} X_{1i}X_{2i} + \hat{\beta}_2\sum_{i=1}^{n} X_{2i}{}^2$

Como tenemos un sistema de 3 ecuaciones con 3 variables ($\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$):

$$
\begin{cases}
\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n} X_{1i} + \hat{\beta}_2\sum_{i=1}^{n} X_{2i} \\
\sum_{i=1}^{n} X_{1i}Y_i = \hat{\beta}_0\sum_{i=1}^{n} X_{1i} + \hat{\beta}_1\sum_{i=1}^{n} X_{1i}{}^2 + \hat{\beta}_2\sum_{i=1}^{n} X_{1i}X_{2i} \\
\sum_{i=1}^{n} X_{2i}Y_i = \hat{\beta}_0\sum_{i=1}^{n} X_{2i} + \hat{\beta}_1\sum_{i=1}^{n} X_{1i}X_{2i} + \hat{\beta}_2\sum_{i=1}^{n} X_{2i}{}^2
\end{cases}
$$

Este sistema de ecuaciones se puede resolver de manera simultánea, utilizando cualquiera de los siguientes métodos: a) eliminación gaussiana, b) sustitución, c) igualación o, d) matricial. Para ilustrar este procedimiento vea el ejemplo siguiente:

Ejemplo 115

A un productor de concentrado para cerdos le gustaría determinar qué relación existe entre la edad de un cerdo cuando empieza a recibir un complemento alimenticio de reciente creación, el peso inicial del animal y la cantidad de peso que aumenta en un período de una semana con el complemento alimenticio. La siguiente información es resultado de un estudio hecho sobre ocho lechones:

Número de lechón	X_1 Peso inicial (libras)	X_2 Edad inicial (semanas)	Y Incremento peso (libras)
1	39	8	7
2	52	6	6
3	49	7	8
4	46	12	10
5	61	9	9
6	35	6	5
7	25	7	3
8	55	4	4

Estime la ecuación de mínimos cuadrados que mejor describa la relación entre Y y las Xs.

Procedimiento:

X_{1i}	X_{2i}	Y_i	X_{1i}^2	X_{2i}^2	$X_{1i}X_{2i}$	$X_{1i}Y_i$	$X_{2i}Y_i$	Y_i^2
39	8	7	1521	64	312	273	56	49
52	6	6	2704	36	312	312	36	36
49	7	8	2401	49	343	392	56	64
46	12	10	2116	144	552	460	120	100
61	9	9	3721	81	549	549	81	81
35	6	5	1225	36	210	175	30	25
5	7	3	625	49	175	75	21	9
55	4	4	3025	16	220	220	16	16
Sumatorias								
362	59	52	17338	475	2673	2456	416	380

Con estos datos formamos el siguiente sistema de ecuaciones:

$$\begin{cases} 8\,\hat{\beta}_0 + 362\,\hat{\beta}_1 + 59\,\hat{\beta}_2 = 52 & \textit{Ecuación } 1 \\ 362\,\hat{\beta}_0 + 17338\,\hat{\beta}_1 + 2673\,\hat{\beta}_2 = 2456 & \textit{Ecuación } 2 \\ 59\,\hat{\beta}_0 + 2673\,\hat{\beta}_1 + 475\,\hat{\beta}_2 = 416 & \textit{Ecuación } 3 \end{cases}$$

Para iniciar, las ecuaciones 1 y 3 se dejan en función de $\hat{\beta}_0$

$$8\,\hat{\beta}_0 = 52 - 362\,\hat{\beta}_1 - 59\,\hat{\beta}_2 \qquad 8362\,\hat{\beta}_0 + 17338\,\hat{\beta}_1 + 2673\,\hat{\beta}_2 = 2456$$

$$\hat{\beta}_0 = \frac{52 - 362\,\hat{\beta}_1 - 59\,\hat{\beta}_2}{8} \qquad \hat{\beta}_0 = \frac{2456 - 17338\,\hat{\beta}_1 - 2673\,\hat{\beta}_2}{362}$$

Igualando estas dos ecuaciones, tenemos que:

$$\frac{52-362\,\hat{\beta}_1-59\,\hat{\beta}_2}{8}=\frac{2456-17338\,\hat{\beta}_1-2673\,\hat{\beta}_2}{362}$$

$$362\,(52-362\,\hat{\beta}_1-59\,\hat{\beta}_2)=8\,(2456-17338\,\hat{\beta}_1-2673\,\hat{\beta}_2)$$

$$\hat{\beta}_1=\frac{824-26\,\hat{\beta}_2}{7660}$$

Ahora en la Ecuación 3 se sustituyen $\hat{\beta}_0$ (cualquiera de las dos definidas anteriormente) y $\hat{\beta}_1$:

$$59\,\hat{\beta}_0+2673\,\hat{\beta}_1+475\,\hat{\beta}_2=416$$

$$59\left(\frac{52-362\,\hat{\beta}_1-59\,\hat{\beta}_2}{8}\right)+2673\left(\frac{824-26\,\hat{\beta}_2}{7660}\right)+475\,\hat{\beta}_2=416$$

$$671.04-2669.75\,\hat{\beta}_1+30.802\,\hat{\beta}_2=416$$

En esta última ecuación sustituimos $\hat{\beta}_1$:

$$59\left(\frac{52-362\,\hat{\beta}_1-59\,\hat{\beta}_2}{8}\right)+2673\left(\frac{824-26\,\hat{\beta}_2}{7660}\right)+475\,\hat{\beta}_2=416$$

$$671.04-2669.75\left(\frac{824-26\,\hat{\beta}_2}{7660}\right)+30.802\,\hat{\beta}_2=416$$

$$383.85+39.864\,\hat{\beta}_2=416 \qquad \hat{\beta}_2=\frac{32.15}{39.864}=0.8065$$

Ahora podemos sustituir el valor de $\hat{\beta}_2$ en la ecuación: $\hat{\beta}_1=\dfrac{824-26\,\hat{\beta}_2}{7660}$, quedando:

$$\hat{\beta}_1=\frac{824-26\,(0.8065)}{7660}=0.1048$$

Finalmente, obtenemos el valor de $\hat{\beta}_0$ usando cualquiera de las siguientes ecuaciones:

$\hat{\beta}_0=\dfrac{52-362\,\hat{\beta}_1-59\,\hat{\beta}_2}{8}$ o $\hat{\beta}_0=\dfrac{2456-17338\,\hat{\beta}_1-2673\,\hat{\beta}_2}{362}$; por ejemplo:

$$\hat{\beta}_0=\frac{52-362\,(0.1048)-59\,(0.8065)}{8}=\frac{-33.5211}{8}=-4.1901$$

8.4 SUPUESTOS ACERCA DEL TÉRMINO DE ERROR ε EN EL MODELO

1. El error ε es una variable aleatoria cuyo valor medio o esperado es cero; esto es, E(ε)=0

2. La varianza de ε se representa por σ^2 y es igual para todos los valores de las variables independientes X_1, X_2, . . . , $X_{k.}$

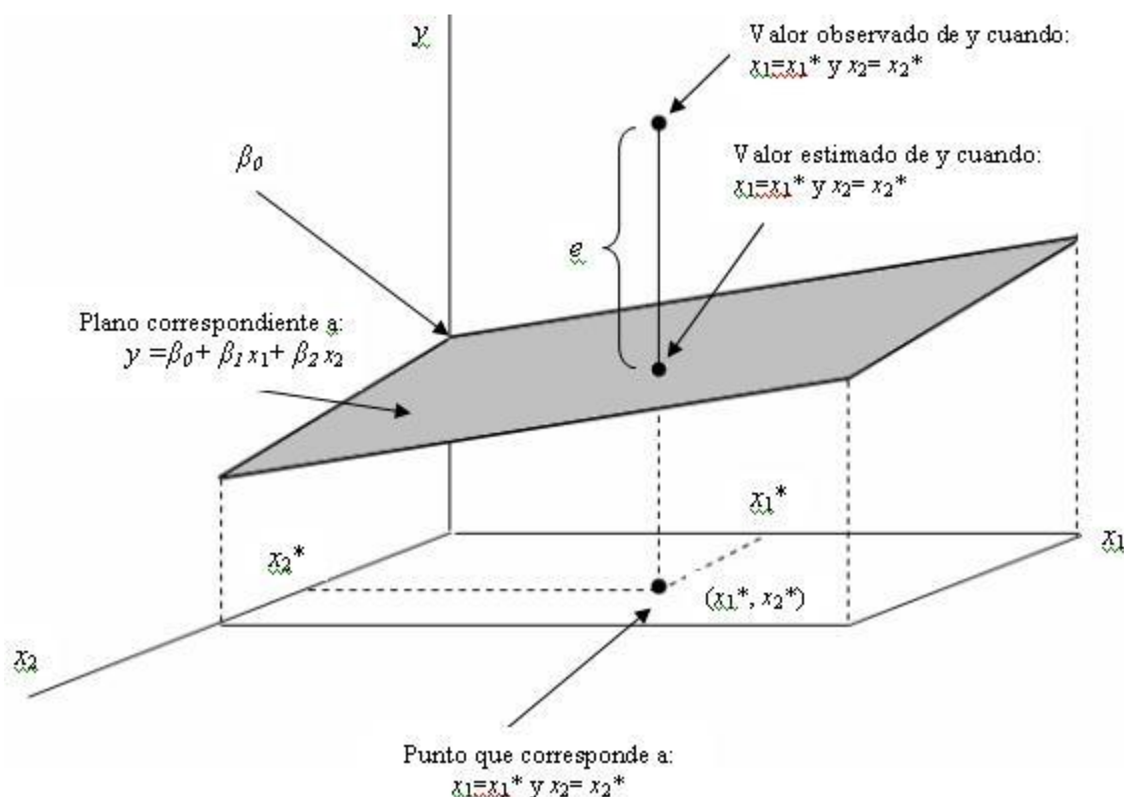
3. Los valores de ε son independientes.

4. El error ε es una variable aleatoria con distribución normal, que refleja la diferencia entre el valor de *y* y el valor esperado de *y*, de acuerdo con $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_k$.

8.5 REPRESENTACIÓN GRÁFICA DE LA ECUACIÓN DE REGRESIÓN LINEAL MÚLTIPLE, CONSIDERANDO DOS VARIABLES INDEPENDIENTES.

Para tener una idea más clara de la relación que expresa la ecuación de regresión lineal múltiple con dos variables independientes: $E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, se presenta a continuación su representación gráfica.

La gráfica de esta ecuación es un plano en el espacio tridimensional. La Figura 1 es esa gráfica con X_1 y X_2 en los ejes horizontales, y *y* en el eje vertical. Observe que ε se muestra como la diferencia entre el valor real de *y* y el valor esperado de *y*, que es E(y), cuando $X_1 = X_1^*$ y $X_2 = X_2^*$.



8.6 EVALUACIÓN DE LA SIGNIFICANCIA DE LA RELACIÓN ENTRE LA VARIABLE DEPENDIENTE Y LAS VARIABLES EXPLICATIVAS (INDEPENDIENTES).

La prueba de F se utiliza para determinar si hay una relación significativa entre la variable dependiente y el conjunto de todas las variables independientes. En estas condiciones, se le llama prueba de significancia global. La hipótesis para la prueba de F implican los parámetros del modelo de regresión múltiple:

$Ho: \beta_1 = \beta_2 = \ldots = \beta_p = 0$ (Y no depende de las X_i)

Ha: al menos un $\beta_i \neq 0$ (Y depende de al menos una de las X_i)

Si se rechaza Ho se tendrá suficiente evidencia estadística para concluir que uno o más de los parámetros no es igual a cero, y que la relación lineal entre *y* y el conjunto de variables independientes $X_1, X_2, \ldots, X_k$ es significativa. El estadístico de prueba F, al igual que en la regresión lineal simple se calcula así:

$$F = \frac{CM_{\text{Regresión}}}{CM_{\text{Residuo}}}$$

Regla de rechazo: se rechaza Ho sí $F > F\alpha$, en donde $F\alpha$ se basa en la distribución de F con *p* grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador.

A continuación se presentan las ecuaciones para realizar el análisis de varianza

Fuentes de variación	**Grados de libertad**	**Suma de Cuadrados (SC)**	**Cuadrados Medios**
Regresión	p–1	$SC_{Reg} = b_0 \sum_{i=1}^{n} Y_i + b_0 \sum_{i=1}^{n} X_{1i} Y_i + b_2 \sum_{i=1}^{n} X_{2i} Y_i - n\overline{Y}^2$	$\frac{SC\ Regresión}{GL\ Regresión}$
Residuo	n–p	$SC_{Res} = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_0 \sum_{i=1}^{n} X_{1i} Y_i - b_2 \sum_{i=1}^{n} X_{2i} Y_i$	$\frac{SC\ Resíduo}{GL\ Resíduo}$
Total	n–1	$SC_{Total} = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$	

Los resultados del Análisis de Varianza obtenidos con Infostat para el ejemplo 114 se presentan a continuación:

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Incremento peso	8	0.88	0.83	2.87	26.93	27.25

Coeficientes de regresión y estadísticos asociados

Coeficientes	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
constante	-4.19	1.89	-9.05	0.66	-2.22	0.0771
Peso inicial	0.10	0.03	0.02	0.19	3.25	0.0228
Edad inicial	0.81	0.16	0.40	1.21	5.10	0.0038

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	37.01	2	18.50	18.54	0.0049
Error	4.99	5	1.00		
Total	42.00	7			

Tal como se observa en la tabla de resumen del Análisis de Varianza, se rechaza la hipótesis nula (valor p = 0.0049) y se concluye que el modelo de regresión lineal múltiple es significativo para interpretar la relación entre incremento de peso y peso inicial + edad inicial de los cerdos.

8.7 INFERENCIAS RELACIONADAS A LOS COEFICIENTES DE REGRESIÓN DE LA POBLACIÓN.

Posteriormente, se recomienda efectuar una prueba de t para determinar el significado de cada uno de los parámetros individuales.

a) Hipótesis

$$Ho: \beta_i = 0 \text{ (X}_i\text{ no es una variable explicativa significativa)}$$

$$Ha: \beta_i \neq 0 \text{ (X}_i\text{ es una variable explicativa significativa)}$$

b) Estadístico de prueba

$$t = \frac{b_i}{S_{b_i}},$$

Siendo que, S_{bi} es el estimado de la desviación estándar de b_i (error típico del parámetro i).

c) Regla de decisión:

Se rechaza Ho si $t < -t_{\alpha/2}$ o si $t > t_{\alpha/2}$, en donde $t_{\alpha/2}$, se basa en una distribución t con n–p–1 grados de libertad.

d) Conclusión

Respecto a β_1 y β_2 se observa que tienen significancia estadística (*)

8.8 EVALUACIÓN DEL AJUSTE DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE: COEFICIENTE DE DETERMINACIÓN (R^2)

El coeficiente de determinación múltiple (R^2) puede interpretarse como la proporción de la variabilidad de la variable dependiente que se puede explicar con la ecuación de regresión múltiple. En consecuencia, cuando se multiplica por 100, se interpreta como la variación porcentual de *y* que se explica con la ecuación de regresión.

Precaución: En general, R^2 aumenta siempre a medida que se agregan variables independientes al modelo. Por esta razón, algunos investigadores prefieren ajustar el R^2 de acuerdo con la cantidad de variables independientes, para evitar una sobreestimación del impacto de agregar una variable independiente sobre la cantidad de variabilidad que explica la ecuación de regresión.

Si *n* representa la cantidad de observaciones y *p* la cantidad de variables independientes, el coeficiente de determinación múltiple ajustado se calcula con la siguiente ecuación:

$$R^2{}_a = 1 - \left[(1 - R^2) \frac{n-1}{n-p-1} \right]$$

Ejemplo 116

Butler Trucking Company, es una empresa independiente de transportes de productos agrícolas. Una gran parte del negocio de Butler tiene que ver con la entregas. Para poder contar con mejores programas de trabajo se desea estimar el tiempo diario total que viajan sus operadores.

Los directivos consideran que ese tiempo total diario (horas) se relaciona estrechamente con la cantidad de millas recorridas para hacer las entregas diarias y con la cantidad de entregas.

Una muestra aleatoria simple de 10 entregas, suministró los siguientes datos:

Recorrido	X_1= Millas recorridas	X_2 = Cantidad de entregas	y = tiempo de recorrido (horas)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Resultados:

	Coeficientes
Intercepto	–0.8687
X1	0.0611
X2	0.9234

La ecuación estimada de regresión es: $\hat{y}$ = –0.8687 + 0.0611 X_1 +0.9234 X_2. La interpretación de los coeficientes se brinda a continuación:

b_1 = 0.0611, indica que 0.0611 horas es un estimado del aumento esperado en tiempo de viaje que corresponde a una milla de distancia recorrida cuando la cantidad de entregas se mantiene constante.

b_2 = 0.9234, un estimado del aumento esperado en el tiempo de viaje que corresponde a un aumento de una entrega, cuando se mantiene constante la cantidad de millas recorridas, es de 0.9234 horas.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.95
Coeficiente de determinación R^2	0.90
R^2 ajustado	0.88
Error estándar	0.57
Observaciones	10

En este ejemplo, el 90.4% de la variabilidad en el tiempo de viaje, *y*, se explica con la ecuación de regresión múltiple, con las millas recorridas y la cantidad de entregas como variables independientes.

Luego se efectuó una prueba de t para determinar el significado de cada uno de los parámetros individuales. Esta prueba se muestra a continuación:

	Coeficientes	Error típico	Estadístico t	Prob.	Inferior 95%	Superior 95%
Intercepto	-0.8687	0.9515	–0.9129	–2.36	–3.1188	1.3814
X1	0.0611	0.0099	6.1824	2.36 *	0.0378	0.0845
X2	0.9234	0.2211	4.1763	2.36 *	0.4006	1.4463

De acuerdo con estos resultados, las dos variables independientes ingresan al modelo de regresión lineal múltiple.

8.9 ERROR ESTÁNDAR DE LA ESTIMACIÓN DE LA REGRESIÓN MÚLTIPLE

Luego de determinar la ecuación de regresión múltiple, es necesario obtener una medida de la dispersión alrededor del plano de regresión múltiple.

En la regresión simple, la estimación se hace más precisa conforme el grado de dispersión alrededor de la recta de regresión se hace más pequeño. Lo mismo se aplica a los puntos de muestra que se encuentran alrededor del plano de regresión múltiple. Para medir esta variación se utiliza de nuevo la medida conocida como: Error estándar de la estimación (Se), y que se calcula con la siguiente ecuación:

$$Se = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-k-1}},$$

en la que:

y_i	=	i-ésimo valor de muestra de la variable dependiente.
$\hat{y}_i$	=	i-ésimo valor estimado a partir de la ecuación de regresión.
n	=	número de puntos de datos de la muestra.
k	=	número de variables independientes.

Así como en la regresión lineal simple, se puede utilizar el error estándar de la estimación y la distribución t de Student para formar un intervalo de (1–α)% de confianza aproximado alrededor de los valores estimados $\hat{y}_i$. Los límites del intervalo de (1–α)% de confianza se calculan así:

$$IC[\hat{y}_i]_{(1-\alpha)} = \hat{y}_i \pm t_{(\alpha/2)} \times Se$$

t se calcula con $n - k - 1$ grados de libertad.

Observación (i)	y_i	*y estimada* $\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	9.3	8.94	0.36	0.131
2	4.8	4.96	-0.16	0.025
3	8.9	8.94	-0.04	0.001
4	6.5	7.09	-0.59	0.350
5	4.2	4.03	0.17	0.027
6	6.2	5.87	0.33	0.110
7	7.4	6.49	0.91	0.834
8	6	6.80	-0.80	0.638
9	7.6	7.40	0.20	0.039
10	6.1	6.48	-0.38	0.145
		Sumatoria	0.00	2.30
		Valor de t	2.36	

Resultados:

$$Se = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-k-1}} = \sqrt{\frac{2.30}{10-2-1}} = 0.57$$

y estimada	Límites	
$\hat{y}_i$	Superior	Inferior
8.94	10.29	7.59
4.96	6.31	3.61
8.94	10.29	7.59
7.09	8.44	5.74
4.03	5.38	2.69
5.87	7.22	4.52
6.49	7.83	5.14
6.80	8.15	5.45
7.40	8.75	6.06
6.48	7.83	5.13

8.10 SELECCIÓN DE VARIABLES

Cuando se desarrolla un modelo de regresión múltiple, el objetivo es utilizar solamente aquellas variables explicativas que sirvan para prever el valor de una variable dependiente. Si una variable explicativa no sirve para hacer esa previsión, ella puede ser excluida del modelo de regresión múltiple, y un modelo con una cantidad menor de variables explicativas puede ser empleado en su lugar.

Un método utilizado para determinar la contribución de una variable explicativa es llamado: Criterio de la prueba de F parcial. Esta prueba incluye la determinación de la contribución, por parte de cada variable explicativa, para el modelo de la suma de cuadrados debida a la regresión, después que todas las otras variables explicativas han sido incluidas en el modelo. La nueva variable explicativa es incluida solamente si genera una mejora significativa en el modelo. Retomando el ejemplo de la empresa Butler Trucking. Sí X_1 representa la cantidad de millas recorridas y es la única variable independiente, se tiene la siguiente ecuación de regresión lineal simple:

$$\hat{y} = 1.27 + 0.0678\, X_1$$

y una suma de cuadrados del error (SQe) = 8.029. Cuando se agrega X_2, la cantidad de entregas, como segunda variable independiente, se obtiene la siguiente ecuación de regresión:

$\hat{y} = -0.8687 + 0.0611\, X_1 + 0.9234\, X_2$ y una SQe = 2.229.

Al agregar X_2 se observa una reducción de la suma de cuadrados del error, ahora la pregunta es ¿Será que al agregar X_2 se obtiene una reducción significativa de SQe?
Se utilizará la notación SQe(X_1) para representar la suma de cuadrados de error cuando X_1 es la única variable independiente en el modelo, y SQe(X_1,X_2) para representar la suma de cuadrados de error cuando en el modelo están X_1 y X_2 a la vez. Por consiguiente, la reducción de la SQe que se obtiene al agregar X_2 al modelo que solo tiene X_1 es:

$$SQe(X_1) - SQe(X_1, X_2) = 8.029 - 2.299 = 5.730$$

Para determinar si esta reducción es significativa, se hace una prueba de F. Las hipótesis para evaluar la contribución de X_2 para el modelo son:

Ho: La variable X_2 no mejora significativamente el modelo, una vez, que X_1 ha sido incluida.
Ha: La variable X_2 mejora significativamente el modelo, una vez, que X_1 ha sido incluida.

El numerador de la estadística F es la reducción de la SQe dividida entre la cantidad agregada de variables al modelo original. En este caso solo se agregó una variable X_2, entonces el numerador de la estadística F es:

$$\frac{\text{SQe}(X_1) + \text{SQe}(X_1,X_2)}{1} = 5.73$$

El resultado es una medida de la reducción de SQe por cada variable agregada al modelo. El denominador de la estadística F es el error promedio al cuadrado para el modelo que tiene todas las variables independientes. Para el ejemplo que se viene trabajando, esto corresponde a que el modelo contenga X_1,X_2 a la vez, así k = 2 y

$$\frac{\frac{\text{SQe}(X_1) - \text{SQe}(X_1,X_2)}{1}}{\frac{\text{SQe}(X_1,X_2)}{n-k-1}}$$

Para esta prueba F, los grados de libertad del numerador son iguales a la cantidad de variables agregadas al modelo, y los del denominador son iguales a $n - k - 1$. Para el problema de Butler Trucking se obtiene:

$$\frac{\frac{5.70}{1}}{\frac{2.229}{7}} = \frac{5.730}{0.3284} = 17.45$$

Este valor de F (17.45), se compara con un valor crítico de F (1,7,0.05) = 5.59. Como el valor de F > Fcrítico, se rechaza la hipótesis nula y se concluye que al agregar X_2 al modelo donde solo estaba X_1 se obtiene una reducción significativa de la suma de cuadrados del error.

CARACTERÍSTICAS DE ALGUNOS PROCEDIMIENTOS DE SELECCIÓN DE VARIABLES

Es común que el investigador tenga interés en seleccionar un modelo que mejor represente el fenómeno en estudio, es decir, aquel que mejor se ajusta a los datos con que cuenta. Para ello el programa SAS presenta varias opciones para realizar la selección en forma automática.

A continuación se describen brevemente tres procedimientos, en los cuales se presenta el resumen de resultados utilizando los datos del ejemplo anterior.

1. Selección hacia adelante o Ascendente (Forward Selection)

a) Encuentra el mejor modelo con una variable. Es decir, selecciona la X con la mayor correlación con Y. En este ejemplo, *Perlac* tiene con *Leche* tiene la r más alta.

b) Calcula la correlación de Y con las otras variables (X´s), manteniendo Xi constante (en este caso Xi=perlac) e incluye la variable con la mayor correlación parcial, Xj (digamos Xj = *Pesec*).

c) Selecciona variables con la mayor contribución en la explicación de Y.

d) A medida que cada variable se incorpora al modelo, los siguientes valores son examinados:

d.1) R^2

d.2) La prueba de F parcial para la variable que recientemente entró al modelo, la cual muestra si la variable ha removido suficiente cantidad de variación en comparación con aquella removida por las variables que entraron previamente al modelo.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Perlac	1	0.9527	0.9527	1.8169	362.75	<.0001
2	Pesec	2	0.0040	0.9567	2.3187	1.56	0.2285

No other variable met the 0.5000 significance level for entry into the model.

2. Selección hacia atrás o descendente (Backward Elimination)

a) Se calcula una ecuación de regresión con todas las variables de interés (modelo completo)

b) Se calcula la prueba de F parcial para cada variable que entró al modelo y se considera a esta como si fuera la última variable a entrar a dicho modelo.

c) El valor más bajo de la prueba parcial de F, digamos FL, se compara a un nivel de significancia preseleccionado de F, digamos Fo. Si FL < Fo se remueve la variable XL del modelo, en caso contrario, se adopta el modelo que se ajustó.

d) Ajuste el modelo con la variable XL removida.

e) Continúe hasta que no pueda remover más variables.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Edpar	2	0.0008	0.9567	2.3187	0.32	0.5802
2	Pesec	1	0.0040	0.9527	1.8169	1.56	0.2285

3. Selección por pasos o Progresivo (Stepwise)

a) Igual que Forward: empiece con la matriz de correlación simple e incluya en la regresión la variable X más altamente correlacionada con la respuesta.

b) Usando los coeficientes de correlación parcial como antes, seleccione la variable a incluir en la regresión, es decir, aquella variable X´s con la mayor correlación parcial con la de respuesta.

c) Examine la contribución que hubiera hecho la primer variable al modelo (digamos Xi), si la segunda variable (digamos Xj) hubiera entrado primero, a través de una prueba parcial de F, es

decir, a cada etapa todas las variables son examinadas por su contribución única (F parcial) al modelo y aquellas que no satisfacen un criterio previamente establecido son eliminadas.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Perlac		1	0.9527	0.9527	1.8169	362.75	<.0001

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Criterios para la adición de nuevas variables al modelo

a) R^2 y el cambio en R^2. Note que R^2 siempre aumenta con la adición de nuevas variables, por lo que el número de parámetros siempre debe ser menor que el tamaño de la muestra.

b) Busque que la suma de cuadrados de residuos sea mínima.

c) Analice el coeficiente de variación. Valores grandes indican mayores desviaciones o mayor varianza.

A continuación se presenta el programa y los resultados, cuando se utiliza el programa estadístico SAS (Statistical Analysis System):

```
options nodate nonumber;
data MIAPA;
input X1 X2 Y;
cards;
100     4     9.3
50      3     4.8
100     4     8.9
100     2     6.5
50      2     4.2
80      2     6.2
75      3     7.4
65      4     6.0
90      3     7.6
90      2     6.1
;
proc reg;
model Y=X1 X2/selection=stepwise;
run;
```

Indica el método de selección de variables:
a) Stepwise: paso a paso
b) Backward: paso atrás
c) Forward: paso al frente

The SAS System

```
                                The REG Procedure
                                  Model: MODEL1
                              Dependent Variable: Y

                            Stepwise Selection: Step 1

              Variable X1 Entered: R-Square = 0.6641 and C(p) = 18.4411

                               Analysis of Variance

                                       Sum of          Mean
    Source                   DF       Squares        Square    F Value    Pr > F

    Model                     1      15.87130      15.87130      15.81    0.0041
    Error                     8       8.02870       1.00359
    Corrected Total           9      23.90000
                       Parameter     Standard
          Variable      Estimate        Error   Type II SS  F Value  Pr > F

          Intercept      1.27391      1.40074      0.83007     0.83  0.3897
          X1             0.06783      0.01706     15.87130    15.81  0.0041

                           Bounds on condition number: 1, 1
----------------------------------------------------------------------------------------------------
                            Stepwise Selection: Step 2

               Variable X2 Entered: R-Square = 0.9038 and C(p) = 3.0000
                               Analysis of Variance

                                       Sum of          Mean
    Source                   DF       Squares        Square    F Value    Pr > F

    Model                     2      21.60056      10.80028      32.88    0.0003
    Error                     7       2.29944       0.32849
    Corrected Total           9      23.90000

                       Parameter     Standard
          Variable      Estimate        Error   Type II SS  F Value  Pr > F

          Intercept     -0.86870      0.95155      0.27378     0.83  0.3916
          X1             0.06113      0.00989     12.55563    38.22  0.0005
          X2             0.92343      0.22111      5.72925    17.44  0.0042

                                  The SAS System
                                The REG Procedure
                                  Model: MODEL1
                              Dependent Variable: Y

                            Stepwise Selection: Step 2

                      Bounds on condition number: 1.027, 4.1079
----------------------------------------------------------------------------------------------------

                 All variables left in the model are significant at the 0.1500 level.
                           All variables have been entered into the model.
                                  Summary of Stepwise Selection
          Variable      Variable      Number      Partial      Model
  Step    Entered       Removed       Vars In     R-Square    R-Square     C(p)      F Value    Pr > F

    1     X1                               1       0.6641      0.6641     18.4411      15.81    0.0041
    2     X2                               2       0.2397      0.9038      3.0000      17.44    0.0042
```

8.11 MULTICOLINEALIDAD

Un problema importante en la aplicación de la regresión múltiple involucra la posible multicolinealidad de las variables explicativas (independientes). Este problema se refiere a situaciones en que algunas de las variables explicativas son fuertemente correlacionadas entre sí. En esas situaciones, variables colineares no brindan nuevas informaciones, y se torna difícil separar el efecto de esas variables en la variable dependiente. En esos casos, los valores de los coeficientes de regresión para las variables correlacionadas pueden fluctuar drásticamente, dependiendo de las variables que se estén o no incluidas en el modelo.

LISTA DE EJERCICIOS 13

1. En el siguiente cuadro se presentan los datos referentes a la altura (expresada en metros), el diámetro a la altura del pecho (expresada en centímetros) y el volumen (m^3/árbol) de 30 árboles de una especie forestal.

No. árbol	Altura M	DAP Cm	Volumen m^3/árbol	No. árbol	Altura M	DAP cm	Volumen m^3/árbol
1	15.9	12.4	0.058935	16	20.0	26.4	0.464594
2	16.0	14.5	0.112122	17	20.2	25.6	0.441232
3	16.5	16.1	0.142551	18	20.2	27.0	0.490811
4	16.5	21.8	0.261356	19	20.7	22.3	0.343096
5	17.2	15.2	0.13245	20	20.7	26.7	0.491845
6	17.5	16.8	0.164624	21	20.7	29.0	0.580232
7	18.2	22.0	0.293597	22	20.7	30.7	0.650253
8	18.2	24.5	0.364115	23	21.0	18.6	0.242148
9	18.6	19.0	0.223798	24	21.2	17.7	0.221369
10	18.7	24.2	0.365012	25	22.2	25.0	0.462454
11	18.7	28.8	0.516966	26	22.5	24.9	0.464961
12	19.0	18.2	0.209764	27	22.5	30.0	0.674932
13	19.0	22.6	0.323449	28	22.7	20.2	0.308719
14	19.0	27.6	0.4824	29	23.0	25.3	0.490687
15	19.4	21.0	0.285151	30	23.2	30.0	0.69593

a) Construya un diagrama de dispersión.
b) Estime los parámetros de la ecuación de regresión e interprételos en términos prácticos.
c) Realice el análisis de varianza y concluya.
d) Calcule el coeficiente de determinación e interprételo.
e) Calcule el coeficiente de correlación lineal y pruebe la hipótesis Ho: $\rho = 0$
f) Grafique la función de regresión en el diagrama de dispersión.

2. Rebolledo Robles, H. (1999)***** reporta los datos del experimento 6715, realizado en la región del Plan Puebla, México, en donde se estudió la respuesta del cultivo del maíz a la fertilización nitrogenada. En este experimento se evaluó la respuesta del rendimiento del maíz (expresado en kg/ha) a las aplicaciones de nitrógeno. Las dosis de nitrógeno evaluadas fueron: 0, 57, 114, 171, 227, 284 y 342 (kg/ha). Las medias de las tres repeticiones se presentan a continuación:

Dosis N	Rendimiento
0	2129.00
57	3355.33
114	3420.83
171	3277.22

Dosis N	Rendimiento
227	3206.50
284	3093.67
342	3015.33

Evalúe el ajuste de un modelo de regresión para expresar la relación entre dosis de nitrógeno y rendimiento de maíz. Justifique su respuesta.

Fuente: Rebolledo Robles, H.H. 1999. Estimación de modelos regresión a experimentos de fertilización y obtención de dosis óptimas económicas de insumos agrícolas. México: Centro de Documentación, Departamento de Suelos, Universidad Autónoma de Chapingo. 55 p.

3. Los datos siguientes corresponden a 23 árboles de *Pinus tecunumani* de la zona de San Esteban Olancho (Honduras), a los que se les midió el DAP (cm), la altura total (m) y diámetros en diferentes secciones, para obtener el volumen (m^3). Estos árboles se toman sólo como ejemplo para mostrar el procedimiento para construir una tabla de volumen, ya que no es recomendable construirla con tan pocos árboles.

Árbol No.	DAP (D) (cm)	Altura (H) Total (m)	Volumen (V) (m^3)
1	36	28	0.861
2	40	29	1.245
3	42	30	1.412
4	43	25	1.339
5	43	28	1.225
6	43	25	1.117
7	43	30	1.464
8	44	25	0.930
9	44	29	1.321
10	45	22	1.003
11	46	25	1.263
12	46	26	1.175
13	46	27	1.254
14	46	32	1.450
15	48	29	1.537
16	51	32	1.612
17	52	22	1.216
18	52	23	1.132
19	53	25	1.636
20	58	30	2.331
21	59	25	1.522
22	65	32	1.926
23	66	28	2.175

Evalúe el ajuste de los siguientes modelos:

a) $V = b_0 + b_1 D^2 HLn$
b) $V = b_0 + b_1 \ln D + b_2 \ln H$

Fuente: Ferreira Rojas, O. **1994**. Manual de inventarios forestales. 2a. Ed. Siguatepeque: ESNACIFOR. 104 p.

4. Un economista está interesado en predecir la demanda anual de cierto producto, utilizando las siguientes variables independientes:
PRECIO = precio del producto (en dólares)
INGRESO = ingreso del consumidor (en dólares)
SUB = precio de un bien sustituto (en dólares)

Nota: un bien sustituto es aquel que puede suplir a otro bien. Por ejemplo, la margarina es un bien sustituto de la mantequilla.

Se recolectaron datos correspondientes al período 1982 – 1996:

Año	Demanda	Precio ($)	Ingreso ($)	Sub ($)
1982	40	9	400	10
1983	45	8	500	14
1984	50	9	600	12
1985	55	8	700	13
1986	60	7	800	11
1987	70	6	900	15
1988	65	6	1000	16
1989	65	8	1100	17
1990	75	5	1200	22
1991	75	5	1300	19
1992	80	5	1400	20
1993	100	3	1500	23
1994	90	4	1600	18
1995	95	3	1700	24
1996	85	4	1800	21

a) Encuentre la ecuación de regresión de mejor ajuste para estos datos.
b) Evalúe el modelo de regresión lineal múltiple.
c) Según la ecuación de regresión obtenida, ¿qué valor de demanda predeciría si el precio de los productos fue de $6, el ingreso del consumidor de $1,200 y el precio del bien sustituto de $17?

5. Interprete la salida del programa Infostat para los datos del ejemplo 115

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Y	20	0.96	0.95	239963.80	297.40	302.37

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	-1186.09	1319.39	-3983.08	1610.89	-0.90	0.3820
X1	9.39	16.64	-25.88	44.67	0.56	0.5802
X2	-7.87	6.85	-22.38	6.64	-1.15	0.2673
X3	27.14	6.88	12.55	41.74	3.94	0.0012

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	45972569.08	3	15324189.69	120.29	<0.0001
Error	2038285.92	16	127392.87		
Total	48010855.00	19			

Selección Stepwise.
Máximo p-valor para entrar: 0.15
Máximo p-valor para retener: 0.15
Número original de regresoras: 3,regresoras retenidas en el modelo 1

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	-2716.38	477.61	-3719.79	-1712.96	-5.69	<0.0001
X3	32.94	1.73	29.30	36.57	19.05	<0.0001

Error cuadrático medio: 126096.963732

BIBLIOGRAFIA

1. Anderson, D.; Sweeney, D.; Williams, T. **2008**. Estadística para Administración y Economía. 10ª Ed. México D.F.: Cengage Learning Editores 1061 p.

2. Andrade, D.; Ogliari, P. **2010**. Estatística para as Ciências Agrárias e Biológicas. 2ª Ed. Florianópolis: Editora da UFSC. 467 p.

3. Balzarini, M. et al. **2011**. Introducción a la Bioestadística. Aplicaciones con Infostat en Agronomía. Universidad Nacional de Córdoba (Argentina): Brujas. 400 p.

4. Balzarini, M. et al. **2015**. Estadística y Biometría Ilustraciones del Uso de InfoStat en Problemas de Agronomía. 2ª. Ed. Universidad Nacional de Córdoba (Argentina): Brujas. 390 p.

5. Batista, J. **1997**. Notas para acompanhar as aulas da disciplina LME - 216 Introdução à Bioestatística Florestal. Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Ciências Florestais. Piracicaba, SP (Brasil). Disponibles en: http://cmq.esalq.usp.br/tutoriais/lme216.pdf

6. Bussab, W.; Morettin, P. **2002**. Estatística básica. 5ª Ed. São Paulo: Saraiva. 526 p.

7. Di Rienzo, JA. et al. **2009**. Estadística para las Ciencias Agropecuarias. 7ª Ed. Córdoba: Brujas. 356 p.

8. Di Rienzo, J.A.; Casanoves, F.; Balzarini, M.G.; González, L.; Tablada, M.; Robledo, C.W. InfoStat versión **2015**. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar

9. Jayaraman, K. **2000**. A statistical manual for forestry research. Bangkok: FORSPA-FAO publication. 240 p.

9. Levin, R.; Rubin, D. **2004**. Estadística para Administración y Economía. 7ª. Ed. México: Pearson Prentice Hall. 952 p.

10. López Velásquez, P.A. **2003**. Estudio del crecimiento y rendimiento de *Pinus pseudostrobus* Lindley, en bosques naturales de los departamentos de Chimaltenango y Sololá. Tesis Ing. Agr. Universidad de San Carlos de Guatemala, Facultad de Agronomía. 101 p.

11. Magalhães, M.N. ; Lima, A. C. De **2001**. Noções de probabilidade e Estatística. 3ª Ed. São Paulo: IME-USP. 392 p.

12. Mendenhall, W. **1991**. Introducción a la probabilidad y a la Estadística. Wadsworth International Iberoamérica.

13. Mendenhall, W.; Scheaffer, R.; Wackerly, D. **2010**. Estadística Matemática con aplicaciones. 7ª. Ed. México, D.F.: Cengage Learning. 939 p.

14. Rustom, A. **2012**. Estadística descriptiva, probabilidad e inferencia, una visión conceptual. Santiago de Chile: Universidad de Chile, Facultad de Ciencias Agronómicas. 197 p. Versión digital disponible en: www.agren.cl/estadistica

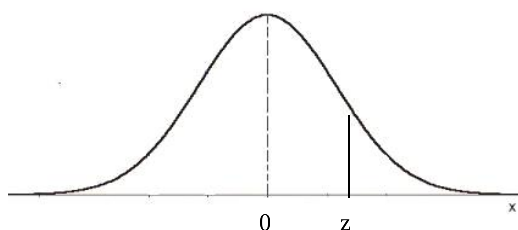
15. Trejos, J.; Moya, J. **1998**. Introducción a la estadística descriptiva. San José, C.R.: Sello Latino. 200 p.

16. Zar, J.H. **1999**. Bioestatistical Analysis. 4ª Ed. New Jersey: Prentice Hall Inc. 929 p.

17. Zocchi, S.; Leandro, R. **2000**. Notas para acompanhar a disciplina LCE 211 – Estatística Geral –. Universidade de São Paulo, Escola Superior de Agricultura “Luiz de Queiroz”, Departamento de Ciências Exatas. Piracicaba, SP (Brasil).

ANEXOS: TABLAS ESTADÍSTICAS
(Generadas en MS Excel® 2013)

DISTRIBUCIÓN NORMAL



Área bajo la curva Normal de 0 a Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.2737	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998	0.49998	0.49998	0.49998

Tabla t de Student

	0.10	0.05	0.025	0.01	0.005	**Una cola**
gl	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**Dos colas**
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
4	1.533	2.132	2.776	3.747	4.604	
5	1.476	2.015	2.571	3.365	4.032	
6	1.440	1.943	2.447	3.143	3.707	
7	1.415	1.895	2.365	2.998	3.499	
8	1.397	1.860	2.306	2.896	3.355	
9	1.383	1.833	2.262	2.821	3.250	
10	1.372	1.812	2.228	2.764	3.169	
11	1.363	1.796	2.201	2.718	3.106	
12	1.356	1.782	2.179	2.681	3.055	
13	1.350	1.771	2.160	2.650	3.012	
14	1.345	1.761	2.145	2.624	2.977	
15	1.341	1.753	2.131	2.602	2.947	
16	1.337	1.746	2.120	2.583	2.921	
17	1.333	1.740	2.110	2.567	2.898	
18	1.330	1.734	2.101	2.552	2.878	
19	1.328	1.729	2.093	2.539	2.861	
20	1.325	1.725	2.086	2.528	2.845	
21	1.323	1.721	2.080	2.518	2.831	
22	1.321	1.717	2.074	2.508	2.819	
23	1.319	1.714	2.069	2.500	2.807	
24	1.318	1.711	2.064	2.492	2.797	
25	1.316	1.708	2.060	2.485	2.787	
26	1.315	1.706	2.056	2.479	2.779	
27	1.314	1.703	2.052	2.473	2.771	
28	1.313	1.701	2.048	2.467	2.763	
29	1.311	1.699	2.045	2.462	2.756	
30	1.310	1.697	2.042	2.457	2.750	
31	1.309	1.696	2.040	2.453	2.744	
32	1.309	1.694	2.037	2.449	2.738	
33	1.308	1.692	2.035	2.445	2.733	
34	1.307	1.691	2.032	2.441	2.728	
35	1.306	1.690	2.030	2.438	2.724	
36	1.306	1.688	2.028	2.434	2.719	
37	1.305	1.687	2.026	2.431	2.715	
38	1.304	1.686	2.024	2.429	2.712	
39	1.304	1.685	2.023	2.426	2.708	
40	1.303	1.684	2.021	2.423	2.704	
50	1.299	1.676	2.009	2.403	2.678	
60	1.296	1.671	2.000	2.390	2.660	
70	1.294	1.667	1.994	2.381	2.648	
80	1.292	1.664	1.990	2.374	2.639	
90	1.291	1.662	1.987	2.368	2.632	
100	1.290	1.660	1.984	2.364	2.626	

Área correspondiente al extremo derecho de una distribución Jí-cuadrada.

gl	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.00098	0.00393	0.01579	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	3.48954	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84391	15.37916	17.29189	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	40.25602	43.77297	46.97924	50.89218	53.67196
40	20.70654	22.16426	24.43304	26.50930	29.05052	51.80506	55.75848	59.34171	63.69074	66.76596
50	27.99075	29.70668	32.35736	34.76425	37.68865	63.16712	67.50481	71.42020	76.15389	79.48998
60	35.53449	37.48485	40.48175	43.18796	46.45889	74.39701	79.08194	83.29768	88.37942	91.95170
70	43.27518	45.44172	48.75757	51.73928	55.32894	85.52704	90.53123	95.02318	100.42518	104.21490
80	51.17193	53.54008	57.15317	60.39148	64.27785	96.57820	101.87947	106.62857	112.32879	116.32106
90	59.19630	61.75408	65.64662	69.12603	73.29109	107.56501	113.14527	118.13589	124.11632	128.29894
100	67.32756	70.06490	74.22193	77.92947	82.35814	118.49800	124.34211	129.56120	135.80672	140.16949

TABLA F, NIVEL DE SIGNIFICANCIA = 5%

Grados de libertad denominador	Grados de libertad del numerador 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91	244.69	245.36	245.95	246.46	246.92	247.32	247.69	248.01	248.31	248.58	248.83	249.05	249.26
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45	19.45	19.45	19.45	19.45	19.46
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66	8.65	8.65	8.64	8.64	8.63
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80	5.79	5.79	5.78	5.77	5.77
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56	4.55	4.54	4.53	4.53	4.52
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87	3.86	3.86	3.85	3.84	3.83
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44	3.43	3.43	3.42	3.41	3.40
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15	3.14	3.13	3.12	3.12	3.11
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94	2.93	2.92	2.91	2.90	2.89
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77	2.76	2.75	2.75	2.74	2.73
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65	2.64	2.63	2.62	2.61	2.60
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54	2.53	2.52	2.51	2.51	2.50
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46	2.45	2.44	2.43	2.42	2.41
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39	2.38	2.37	2.36	2.35	2.34
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33	2.32	2.31	2.30	2.29	2.28
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30	2.29	2.28	2.26	2.25	2.24	2.24	2.23
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23	2.22	2.21	2.20	2.19	2.18
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19	2.18	2.17	2.16	2.15	2.14
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16	2.14	2.13	2.12	2.11	2.11
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12	2.11	2.10	2.09	2.08	2.07
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.11	2.10	2.08	2.07	2.06	2.05	2.05
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	2.11	2.09	2.08	2.06	2.05	2.04	2.02	2.01	2.01	2.00
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.09	2.07	2.05	2.04	2.03	2.01	2.00	1.99	1.98	1.97
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.07	2.05	2.04	2.02	2.01	2.00	1.98	1.97	1.96	1.96